Genus-scale Analysis of Gene Cluster Evolution in Fungi

Kjærbølling, Inge

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

Genus-scale Analysis of Gene Cluster Evolution in Fungi

PhD Thesis

Inge Kjærbølling

Department of Biotechnology and Biomedicine
Technical University of Denmark

August 2018

Supervisors

Professor MSO Mikael R. Andersen
Assistant Professor Tammi Vesth
Professor Uffe H. Mortensen
Professor Thomas O. Larsen

Preface

This thesis serves as a partial fulfillment of the requirements to obtain a PhD degree from the Technical University of Denmark (DTU). The work was carried out under the supervision of Mikael R. Andersen, Tammi Vesth, Thomas O. Larsen, and Uffe H. Mortensen from September 2015 to August 2018. With the changing organization of the biosciences at DTU, the work was initiated at the then Department of Systems Biology and finished at Department of Biotechnology and Biomedicine. The project has been funded by a DTU PhD stipend.

Part of the work presented in section 3.2 was conducted during my external stay at Kanazawa Institute of Technology under the supervision of Prof. Masayuki Machida.

Kgs. Lyngby, August 2018

Inge Kjærbølling

Acknowledgements

After three years of studies, countless hours of work, enormous amounts of coffee, transformations, queries, data overload, laughs, success, failures, life lessons, blood, sweat, and tears I would like to thank the people who have helped me through it all.

First and foremost I would like to express my deepest gratitude to my supervisors Mikael R. Andersen and Tammi Vesth. To Mikael, for introducing me to the DTU scholarship and agreeing to be my supervisor and for your thoughtful inputs, support, great ideas and for providing puppies when needed. To Tammi, for your always constructive feedback, encouragement, support, and supply of good coffee. I have learned a lot from both of you scientifically as well as personally and I could not have wished for a better supervisor team.

I would also like to thank the Network Engineering of Eukaryotic Cell Factories group. Especially, Jane for your positive morning energy and discussion and Sebastian four your great inputs and laughs. We have had a lot of awesome times together especially travelling to conferences. You have both made some excellent work that I have been able to build upon and I truly appreciate our cooperation.

My former office mates Christina and Maria I would like to thank for introducing me to the *Aspergillus* lab and getting me started with fungal molecular biology. In addition, I would like to give thanks to the people working in the *Aspergillus* lab for high spirits, good music and helpfulness. My current office mates Gosia and Katherina I would like to thank for support, survivor attitude and late laughs.

Thomas I. Petersen and Sara Kildgaard I would like to thank for your great work on the chemical analysis of the section *Flavi* species. Moreover, I greatly appreciate the work Sara performed on the chemical analysis analysis of my mutants strains.

I am grateful to professor Masayuki Machida for hosting me during 2 months external research stay at Kanazawa Institute of Technology. I value our fruitful discussions scientifically as well as culturally and appreciate the introduction you gave me to the Japanese cuisine. I would also like to thank my co-supervisors Thomas O. Larsen and Uffe H. Mortensen for their inputs and suggestions when I needed them.

I owe a huge thanks to my friends and family for support and understanding. My grandparents, parents and sisters who show an impeccable faith in me and always give loving support. The biotek10 crew, without whom it would not have been half as much fun studying biotechnology at DTU, always ready for social

gatherings, scientific discussions and brainstorming and providing endless biotek love and great karma.

Finally, I would like to give a very special thanks to Jonas for your love, patience, support, and encouragement through it all.

Abstract

The *Aspergillus* genus is highly diverse and contains more than 300 species. Several of these have high impact on society, beneficial as well as harmful, including the opportunistic pathogen *A. fumigatus*, the food spoiler *A. flavus*, the citric acid and enzyme producer *A. niger*, and the food fermentor *A. oryzae*. *Aspergillus* species are known to produce a high number of secondary metabolites and have been shown to have an even higher genomic potential based on the number of predicted secondary metabolite gene clusters in the genomes. Some secondary metabolites have bioactivities which are useful medically such as antibiotics, immunosupressants, and cholesterol-lowering agents. The *Aspergillus* whole genus sequencing project (of which this project is a part of) aims to produce a genome sequence for a representative strain of each species within the genus. In this thesis, we investigate both the opportunities and challenges in exploring this genus using the vast data created from the genome sequencing project. We focus on secondary metabolism addressing the challenge of how we can identify beneficial bioactive clusters using the diversity and self-resistance mechanisms. Moreover, we investigate the diversity and similarities of species spanning several sections and across the *Flavi* section.

As part of the *Aspergillus* sequencing project, this thesis work will publish 25 genomes. We have investigated the genome characteristics and used the diversity of the genomes to create novel insights and hence knowledge-based hypothesis. To characterize and utilize a bioactive compound industrially, it is important to know which genes are responsible for the biosynthesis. There are many ways of linking gene clusters to compounds and here we present an overview of these by divide them into strategies. Moreover, we have demonstrated the use of some of the strategies based on comparative genomics and retrobiosynthesis to identify putative clusters for, among others, the bioactive compounds novofumigatonin and chlorflavonin, of which the novofumigatonin cluster has been experimentally verified subsequently [1]. Investigating the *Flavi* section, we have identified common traits such as large genomes and a high number of predicted biosynthetic gene clusters, but also differences such as variations in gene cluster families. Examination of the phylogeny have led to questions regarding the evolution in the *A. flavus* clade and domestication of *A. oryzae* and *A. sojae*. The carbohydrate-active enzyme (CAZy) potential is large within the *Flavi* section with a maximum of 663 CAZymes found in *A. parasiticus*.

With the large number of predicted secondary metabolite gene clusters it is a challenge to select the most promising clusters and prioritize the experimental ef-

forts in the quest of novel bioactive compounds. In order to overcome this challenge we have developed a pipeline, FRIGG (Fungal ResIstance Gene-directed Genome mining), using genome sequences to identify putative bioactive gene clusters based on duplicated self-resistance genes. This pipeline thereby provides a means of selecting which clusters to investigate and dramatically shortens the experimental process. Applying the pipeline to 51 *Aspergillus* and *Penicillium* species, we have identified a total of 72 protein families with putative resistance genes found in clusters including the verified resistance gene for fellutamide B. We selected a cluster for experimental investigation and preliminary results indicate it could be producing N-acetyl-glutamine which have shown bioactivity as a psycostimulant and in a complex with aluminium as an antiulcer agent and [2, 3, 4].

In summary, this work has not only contributed to to the *Aspergillus* community with new genome sequences and insights from comparative genomics analysis, but also with strategies to link gene clusters and compounds and a pipeline identifying putative resistance genes and bioactive clusters. In the future, this pipeline can be used as a guide in the quest for novel bioactive compounds, which are desperately needed. The exploration of the newly sequenced genomes has only just started with our generated insights and hypothesis and it will continue to move the field forward gaining many more insights in the time ahead.

Dansk resumé

Aspergillus slægten er meget divers og indeholder mere end 300 arter, hvoraf flere har stor indvirkning på samfundet; gavnlige såvel som skadelige. Dette inkluderer den opportunistiske patogen *A. fumigatus*, den afgrøde ødelæggende *A. flavus*, citronsyre og enzym producenten *A. niger* samt *A. oryzae*, som bruges til i det asiatiske køkken til fermentering. *Aspergillus* arter er kendte for at producere et højt antal sekundære metabolitter, desuden har de udvist et højt genomisk potentiale baseret på antallet af forventede sekundære metabolit genclustre fundet i genomerne. Nogle sekundære metabolitter har bioaktiviteter, som er brugbare medicinsk såsom antibiotika, immunosuppressiver, og kolesterol sænkende stoffer. *Aspergillus* hel-slægts sekventeringsprojektet (hvilket dette projekt er en del af) sigter efter at generere en genomsekvens for repræsentative stammer af hver art i slægten. I denne afhandling, undersøger vi både de muligheder og udfordringer, der er forbundet med at udforske slægten ved brug af den enorme mængde data genereret af sekventeringsprojektet. Vi fokuserer på sekundær metabolisme og adresserer udfordringen af, hvordan vi identificerer bioaktive genclustre ved at bruge diversiteten og selv-resistensmekanismer. Derudover undersøger vi diversiteten og lighederne mellem arter fra forskellige dele af slægten samt på tværs af *Flavi* sektionen.

Som en del af *Aspergillus* sekventeringsprojektet kommer denne afhandling til at publicere 25 genomer. Vi har undersøgt genom-egenskaberne samt brugt genomerne til at få nye indblik og dermed lave videns baserede hypoteser. For at karakterisere og kunne bruge bioaktive stoffer industrielt er det vigtigt at vide hvilke gener, som er ansvarlige for biosyntesen. Der findes mange måder at koble genclustre og stoffer på, vi har skabt et overblik over disse ved at dele dem ind i strategier. Derudover har vi vist brugen af nogle af disse strategier baseret på sammenlignende genom-analyser og retro-biosyntese til at identificere formodede genclustre for blandt andet de bioaktive stoffer novofumigatonin og chlorflavonin, hvor novofumigatonin clusteret efterfølgende er blevet verificeret eksperimentelt [1]. Ved at undersøge sektion *Flavi* har vi identificeret fælles træk så som store genomer og et højt antal forventede biosyntetiske genclustre, men også forskelle som variationer i cluster familier. Undersøgelse af fylogenien har ført til spørgsmål vedrørende udviklingen i *A. flavus* cladden samt domesticering af *A. oryzae* og *A. sojae*. Det kulhydrataktive enzym potentiale er også stort indenfor sektion *Flavi* med 663 enzymer fundet i *A. parasiticus* som det højeste.

Med det enorme antal formodede sekundære metabolit genclustre er det en udfordring at udvælge de mest lovende clustre og prioritere den eksperimentelle

indsats i jagten efter nye bioaktive stoffer. For at overkomme denne udfordring har vi udviklet en pipline, FRIGG (Fungal ResIstance Gene-directed Genome mining), som bruger genomsekvenser og den genetiske diversitet til at identificere mulige bioaktive genclustre baseret på et duplikeret selv-resistensgen. Dette giver dermed en måde at udvælge hvilke genclustre, man skal fokusere på og undersøge, hvilket drastisk forkorter processen. Ved at anvende pipelinen på 51 *Aspergillus* og *Penicillium* arter har vi identificeret 72 protein familier med formodede resistensgener, som er fundet i genclustre inklusiv det verificerede resistensgen for fellutamid B. Vi udvalgte et gencluster til eksperimentel undersøgelse og de foreløbige resultater indikerer, at det kan være N-acetyl-glutamin, som clusteret producerer hvilket har udvist bioaktivitet mod mavesår (i et kompleks med aluminium) og som psykostimulerende stof [2, 3, 4].

Samlet set har dette arbejde ikke alene bidraget til *Aspergillus* forskningsfeltet med nye genomer og indblik fra sammenlignende genom-analyser men også med strategier til at koble genclustre og stoffer samt en pipeline, som identificerer mulige resistensgener og bioaktive clustre. I fremtiden kan den udviklede pipeline blive brugt som guide i jagten efter nye bioaktive stoffer, hvilket der er et stort behov for. Udforskningen af de ny-sekventerede genomer er kun lige begyndt med vores undersøgelser og frembragte hypoteser. Denne ressource vil fortsætte med at flytte forskningsfeltet og være et vigtigt aktiv for fremtidige forskningsresultater og indsigter.

Publications

During the course of this project I have written 4 manuscripts as first author and contributed to 3 other manuscripts. Of these 2 have been published and 5 are currently in preparation.

Papers included in thesis

Linking secondary metabolites to gene clusters through genome sequencing of six diverse *Aspergillus* species

Kjærbølling I., Vesth T.C., Frisvad J.C, Nybo J.L., Sebastian T., Kuo A., Bowyer P., Matsuda Y., Mondo S., Lyhne E.K., Kogle M.E., Clum A., Lipzen A.M., Salamov A.A., Ngan C.Y., Daum C.G., Chiniquy J., Barry K.W., LaButti K.M., Haridas S., Simmons B.A., Magnuson J.K., Mortensen U.H., Larsen T.O., Grigoriev I.V., Baker S.E. & Andersen M.R.

Published January 9, 2018 – Proceedings of the National Academy of Sciences of the United States of America.

Manuscripts included in thesis

Review coupling compounds to genes and vice versa

Kjærbølling, I., Vesth, T., Mortensen, U.H., Larsen, T.O. & Andersen, M.R.

In preparation for submission to Fungal Genetics and Biology

Friends and foes – A comparative genomics study of 23 Aspergillus species from section Flavi

Kjærbølling, I., Vesth, T., Frisvad, J.C., Nybo, J.L., Theobald, S., Kuo, A., Sato, A., Lyhne, E.K., Kogle, M.E., Clum, A., Lipzen, A., Salamov, A., Ngan, C.Y., Daum, C., Chiniquy, J., Barry, K., LaButti, K., Haridas, S., Simmons, B.A., Magnuson, J.K., Mortensen, U.H., Larsen, T.O., Grigoriev, I.V., Machida, M., Baker S.E. & Andersen, M.R.

In preparation for submission to Genome Biology

Resistance Gene Directed Genome Mining pipeline

Kjærbølling, I., Vesth, T., Mortensen, U.H., Larsen, T.O. & Andersen, M.R.

In preparation for submission to Fungal Biology and Biotechnology.

Papers not included in thesis

Novofumigatonin biosynthesis involves a non-heme iron-dependent endoperoxide isomerase for orthoester formation

Matsuda, Y., Bai, T., Phippen, C.B.W., Nødvig, C.S., Kjærbølling, I., Vesth, T.C., Andersen, M.R., Mortensen, U.H., Gotfredsen, C.H., Abe, I & Larsen, T.O., Nature Communications, 3 July 2018, 9, 2587, DOI: 10.1038/s41467-018-04983-2

Published July 3, 2018 – Nature Communications.

Abbreviations

AMA	Autonomous maintenance in *Aspergillus*
BGC	Biosynthetic gene cluster
bp	Base pair
BPC	Base peak chromatogram
CPA	Cyclopiazonic acid
CRISPR	Clustered regularly interspaced short palindromic repeats
DMAT	Dimethylallyl transferase
DNA	Deoxyribonucleic acid
EIC	Extracted ion chromatogram
ESI	Electro spray ion
FAC-MS	Fungal artifical chromosomes - metabolic scoring
FRIGG	Fungal ResIstance Gene directed Genome mining
gDNA	Genomic DNA
HPLC	High-performance liquid chromatography
HR	Homologous recombination
HRMS	High resolution mass spectrometry
HR-PKS	Highly reducing PKS
IMPDH	Inosine-5'-monophosphate dehydrogenase
IS	Insertion site
LB	Luria-Betrani
MM	Minimal media
MPA	Mycophenolic acid
MS	Mass spectrometry
m/z	Mass to charge
NHEJ	Non homologous end joining
NR	Non reducing
NRPS	Nonribosomal peptide synthetase
PCR	Polymerase Chain Reaction
PKS	Polyketide synthase
RT	Retention time
SM	Secondary metabolite
SMGC	Secondary metabolite gene cluster
TC	Terpene cyclases
TF	Transcription factor
TM	Transformation medium
UHPLC-DAD-QTOFMS	Ultra-high performance liquid chromatography – diode array detection – Quadrupole time-of-flight mass spectrometry
USER	Uracil-Specific Excision Reagent
UV	Ultraviolet
YES	Yeast extract sucrose

Contents

Preface i
Acknowledgements ii
Abstract iv
Dansk resumé vi
Publications viii
Papers included in thesis viii
Manuscripts included in thesis viii
Abbreviations x

1 Introduction **1**
1.1 Project motivation 1
1.2 Project aim and outline 3

2 Background **5**
2.1 *Aspergillus* – a genus with a large impact 5
2.2 Secondary metabolism 7
2.3 Self resistance mechanisms 8
2.4 Manuscript I – Review: linking compounds and gene clusters 10

3 Exploring genomic diversity and linking compounds to clusters **35**
3.1 Paper I – Linking compounds to gene clusters through genome sequencing 35
3.2 Manuscript II – Comparative genomics of section *Flavi* 46

4 Resistance gene-directed genome mining **71**
4.1 Manuscript III - Pipeline for resistance gene-directed genome mining . . 71
4.2 Identified and investigated resistance case 91
4.2.1 Introduction 91
4.2.2 Results and discussion 92
4.2.3 Conclusion 101
4.2.4 Methods 102

5 Conclusion **107**

Bibliography **109**

A Supplementary section 3.1 – Paper I **119**

B Supplementary material section 3.2 – Manuscript II **149**

C Supplementary material section 4.1 – Manuscript III **165**

D Supplementary material section 4.2 – Experimental investigation **169**

1 Introduction

1.1 Project motivation

The fungal kingdom is a highly diverse group of organisms growing in most habitats on earth estimated to include up to 5.1 million species which are extremely important in ecosystems as decomposers and essential associates of other organisms [5, 6]. Moreover, fungal species can cause damaging infections in plants as well as animals and humans and thereby pose a serious threat to society. Fungal infections in plants have long been a recognized problem, one that has even changed human history with the late blight that led to the Irish potato famine. Plant fungal infections are still a serious problem recognized as a threat to food security [7, 8]. More recently, fungal infections are also being recognized as a threat to animal health and ecosystems [8] and fungal infections also contribute substantially to human morbidity and mortality, but is often overlooked. The mortality rate of invasive fungal infections is extremely high (often more than 50%) and it is estimated that more than 1.5 million die from fungal infections every year [9]. Invasive fungal infections often affect immunocompromised patients (due to organ transplant or other underlying diseases) and it is an increasing problem [10, 11, 12, 9].

There are few available drugs treating fungal infections; these can be divided into three classes by the mode of action; triazoles and allylamines targeting sterole synthesis, polyenes and echindocandins targeting the cell wall, and pyrimidine analogs (flucytosine) targeting DNA synthesis [13, 9]. The global market for antifungal drugs was valued to $13.1 billion in 2016 and is estimated to increase and reach $16.1 billion by 2021 [14]. Recently emerging resistance to azoles has been seen in *Aspergillus* species [15, 16, 17] and it has been suggested that the increasing resistance in environmental samples is caused by the massive use of azole fungicides for plant protection in agriculture [18, 19].

The current status is thus the following; fungal infection pose a serious threat to food security, there is an increasing number of invasive fungal infections with high mortality rates, few antifungal agents, and a rise of resistance. This situation desperately calls for action.

Ironically many antifungal compounds (such as griseofulvins and echinocan-

dins) are naturally produced by fungi, thus making them an excellent source of novel antifungals. Especially the *Aspergillus* genus is known to produce an extraordinary number of bioactive compounds (eliciting pharmacological or toxicological effects) known as secondary metabolites [20]. Several of these have been used extensively in the medical industry including the antibiotic penicillin [21] and cholesterol-lowering lovastatin [22]. Under standard laboratory conditions only relatively few compounds are produced, but whole genome sequences of *A. nidulans, A. oryzae* and *A. fumigatus* have revealed a much larger potential [23, 24, 25, 26]. As of 2015 a collaboration between JBEI, DTU, and JGI started genome sequencing all the species in the *Aspergillus* genus, the so-called *Aspergillus* whole genus sequencing project. This project has dramatically increased the available amount of data and the potential of finding novel antifungals. In order to exploit this rich natural resource and fully reap the benefits, the bioactive secondary metabolites have to be identified and knowledge of the biosynthetic gene cluster and the target gene have to be obtained.

The objective of this PhD is not only to rationally identify bioactive secondary metabolites such as new antifungals but also to establish the links to the biosynthetic genes and the target. Furthermore we will advance the comparative genomic research of the *Aspergillus* genus to get a deeper understanding of these important species and their potential uses.

1.2 Project aim and outline

This project falls within the three interrelated subjects; secondary metabolites, biosynthetic gene clusters, and targets affected by secondary metabolites, all connected by genome mining. The unique resource of data from the *Aspergillus* genus project as well as publicly available genome sequences was used in this project. The overall aim of this PhD project was to investigate and establish links between these three subjects – secondary metabolites, genes, and targets. To fulfill this goal, four objectives are laid out:

1. Develop methods for linking compounds and clusters using whole genome sequences and comparative genomics.

2. Publish 25 *Aspergillus* genome sequences to support genome-driven research in the *Aspergillus* genus.

3. Describe and investigate the biological and chemical diversity found within *Aspergillus* species using comparative genomics.

4. Identify and verify clusters containing putative resistance genes using genome sequences.

These aims are addressed throughout the thesis which is is divided into chapters. The chapters are outlined here briefly:

Chapter 2 provides background for the project, introducing the genus *Aspergillus*, secondary metabolism, self-resistance mechanisms, and methods of linking compounds to clusters and vice versa which are presented in manuscript I. This chapter gives the background needed for the following chapters in addition to explaining state-of-the-art strategies for linking compounds and clusters hence addressing aim 1.

Chapter 3 is concerned with comparative genomics and the uses of comparative genomics in linking compound to cluster thereby related to both aims 1, 2 and 3. The chapter includes one published paper and manuscript II. The paper compares the genomes of 13 species and uses this to identify putative biosynthetic gene clusters for specific compounds. Furthermore it compares the opportunistic pathogen *A. fumigatus* to the closely related species *A. novofumigatus* to identify allergens and pathogenicity factors. The manuscript compares 23 section *Flavi* species investigating the diversity, the secondary metabolism, and carbohydrate active enzyme potential across the section.

Chapter 4 presents a pipeline developed for resistance gene directed genome mining. It includes manuscript III demonstrating the pipeline and results from a test dataset of 51 *Aspergillus* and *Penicillium* species. It is followed by a section concerning the investigation of a selected cluster with a putative resistance gene. The third and fourth aim is thereby addressed in this chapter.

Chapter 5 summarizes the work and includes a few future perspective.

The **Appendix** includes additional figures, tables, and information from the paper, chapters, and manuscripts and is ordered based on the sections.

2 Background

2.1 *Aspergillus* – a genus with a large impact

The *Aspergillus* genus is a highly diverse group of filamentous fungi containing more than 300 species [27]. The species are widely distributed worldwide and have an enormous impact on society, beneficial as well as harmful [27, 28]. The genus is currently divided into four subgenera (*Aspergillus, Circumdati, Fumigati* and *Nidulantes*) and further into 20 sections [29, 30].

The main harmful effects include human infections by pathogenic species and food spoilage by mycotoxin producers. *A. fumigatus* is an important opportunistic pathogen and the major cause of Aspergillosis which causes severe and often fatal infections in humans [31, 32, 33]. This is a growing problem due to the increasing number of immunocompromised patients [31, 34, 33]. In addition, several members of the *Aspergillus* genus are known as food and feed spoilers infecting crops and producing mycotoxinc [35]. Aflatoxin an important and highly toxic compound is produced by *A. flavus* and *A. parasiticus* which infects corn and peanuts. The toxicological effect depends on the exposure but ranges from acute effects, including rapid death, and chronic outcomes such as liver cancer [36, 35, 37]. Other important toxins produced by members of the *Aspergillus* genus include ochratoxins, fumonisins, patulin, gliotoxin and cyclopiazonic acid [35].

On the beneficial side, *Aspergillus* species are used in industrial production of enzymes and medical compounds as a model organism for microbiology studies and for food fermentation. The *Aspergillus* genus include some major industrial workhorses such as *A. niger* widely used for citric acid production [38, 39, 40] and enzyme production [38, 41, 42]. *A. oryzae* is also extensively used for enzyme production since it naturally secretes a high amount of enzymes [41, 42, 43]. Both *A. niger* and *A. oryzae* have obtained the status of 'Generally regarded as safe' (GRAS) for several processes [38, 44]. Another use of *A. oryzae* along with *A. sojae* is in food fermentation which has a very long history of use in Asian countries for the production of miso (bean curd seasoning), soy sauce, and sake [44, 45, 46]. In addition, *A. nidulans* is used as a classical model organism for developmental studies and gene regulation aiding the understanding of filamentous fungi [47]. Besides the aforementioned toxic compounds, members of *Aspergillus* also produce

many medically useful compounds [20] including the cholesterol-lowering lovastatin produced by *A. terreus* [48] and the antibiotic penicillin produced by *A. nidulans* [49]. Members of the *Aspergillus* genus are known to produce a high number of these beneficial and toxic compounds (also known as natural products) with an average of 5.77 compounds per species which is significantly higher than the sister genus *Penicillium* producing 3.77 compounds per species [50, 35].

The genomics era of *Aspergillus* was accelerated in 2005 when *A. fumigatus, A. nidulans* and *A. oryzae* were whole genome sequenced [25, 23, 24]. Quickly more species followed [51, 52, 53, 54] and investigation of the genomes not only revealed an immense diversity but also a hidden potential of natural products [26, 55]. Predictions based on the genome sequences indicated that the number of potential compounds could be 10 times higher than the compounds detected by UV and MS [26, 50]. Following the whole genome sequencing, comparative genomics studies have provided insights into evolution of the genus, the species concept, genome dynamics and pathogenicity [56, 57, 58, 59]. In addition it became the start of genome mining for novel natural products [60, 55].

To take the genomics era of *Aspergillus* to the next level, the *Aspergillus* whole genus sequencing project was initiated with the aim of sequencing a representative of each species from the entire *Aspergillus* genus comprising more than 300 species [27]. This ambitious project was started in 2013 by a collaboration of researchers from the Technical University of Denmark, Westerdijk Fungal Biodiversity Institute, DoE Joint BioEnergy Institute and Joint Genome Institute. Another fungal sequencing project, the 1000 fungal genomes (http://1000.fungalgenomes.org/home/), has the purpose of capturing as much diversity as possible across the fungal kingdom sequencing representatives from each of the recognized fungal families covering large evolutionary distances. Hence it will provide reference genomes for future studies and help advance diverse research fields of plant-microbe interactions, microbial emission and capture of greenhouse gasses, and environmental metagenomic sequencing. The *Aspergillus* sequencing project has another more focused purpose mapping the similarities and differences of an entire genus. Having many members from the same genus gives the opportunity to identify traits and mapping genotype to phenotype, this is also ideal for genome mining searching for specific features. The *Aspergillus* project will provide an unseen dataset for the fungal research community which will be used to, gain a deeper understanding of the evolution of the genus, obtain insights for optimization of fungal cell factories, and discover novel natural products.

2.2 Secondary metabolism

As mentioned earlier, members of the *Aspergillus* genus are known to produce a wide range of natural products, also known as secondary metabolites (SMs). Secondary metabolites are, unlike primary metabolites, not essential for normal growth or the survival of the producing organism but they play adaptive roles for example by functioning as signaling molecules in ecological interactions or as defence compounds thereby giving the organism survival benefits [61, 26, 62, 49]. The properties that provides the beneficial survival attributes have also found applications in pharmacology which has ensured significant research within the field of secondary metabolism and drug discovery [63].

The genes responsible for producing secondary metabolites are, most commonly, organized in gene clusters on the genome and they are typically co-regulated [64, 65, 62, 55]. These biosynthetic gene clusters most often consist of 1) one or more backbone gene(s) responsible for production of the core structure of the compound, 2) potentially tailoring enzymes decorating or modifying the core structure and 3) possibly other genes not directly involved in the biosynthesis of the compound such as regulators, transporters and self-resistance genes [26, 62, 65]. Although secondary metabolites are highly diverse and chemically complex, they are all derived from a limited number of precursors from primary metabolism [26]. The biosynthetic backbone genes can therefore be categorized based on the precursors they use; polyketide synthases (PKSs), non-ribosomal peptide synthetases (NRPSs), terpene cyclases (TCs) and prenyltransferases/dimethylallyl transferases (DMATs) [65, 66].

The most common backbone genes are PKSs followed by NRPSs. Both are so-called mega-enzymes consisting of several active domains. PKSs assemble simple acyl units using Claisen C–C bond formation which can be followed by various degrees of reduction or methylation. PKSs in fungi can use each domain several times and are therefore called iterative. Depending on the domains and the polyketide produced, the PKSs can be divided into highly reducing (HR) or partially reducing making macrolides and non-reducing (NR) creating aromatic compounds [67, 68, 69]. NRPSs assembles amino acids into small peptides. Unlike the PKSs, the NRPSs in fungi only use each domain once and hence the enzyme is ordered in modules where each module is responsible for the addition of one unit [69, 70]. Examples of polyketide-derived compounds includes the cholesterol lowering compound, lovastatin [22, 71] and cladosporin with bioactivities counting antifungal, antibiotic, plant-growth inhibitory properties, anti-inflammatory responses and antimalarial [72]. Cladosporin is produced by a combination of a HR- and a NR-

PKS. Well-known non-ribosomal peptides include the anti-fungal echinochandin which is produced by a six-module NRPS in among others *Aspergillus rugulosus* [73]. The immunosuppressant drug cyclosporin also shows both antifungal and antiviral activity, it is produced by an 11 module NRPS by *Tolypocladium inflatum* [74]

2.3 Self resistance mechanisms

Fungi can encounter a wide range of environments and have developed ways of avoiding attacks from biotic agents and coping with abiotic agents such as chemicals, fungicides, salt etc. These coping mechanisms are also known as resistance. Here we focus on a specific kind of resistance, resistance to the factors the fungi produce themselves, self-resistance.

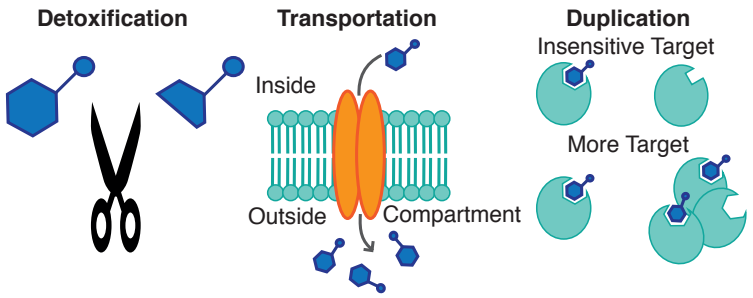


Figure 2.1 Self-resistance mechanisms to secondary metabolites. **Detoxification** – the bioactive toxic form of the secondary metabolite is modified to a nontoxic version within the cell. **Transportation** – the secondary metabolite is exported from the cell or kept in certain enclosed compartments. **Duplication** – the target gene affected by the secondary metabolite is duplicated and either the sheer number of extra target proteins give the resistance, or the duplicated version is insensitive and can tolerate the secondary metabolite.

Many secondary metabolites (SMs) provide certain advantages to the fungi since they have some bioactivity. The bioactivity of the SM is often a result of the SM attacking essential cellular proteins where many of them are also found in the fungi producing that specific SM. In order to avoid the harmful effects and suicide, the producing fungi needs to develop protective mechanisms, self-resistance [63]. These self-resistance mechanisms have been seen to be embedded in the biosynthetic gene cluster encoding the bioactive SM and can be divided

into three main strategies 1) detoxification of the SM 2) transportation of the SM and 3) duplication of the target gene. These mechanisms were reviewed for antibiotic producing bacteria already in the seventies and eighties [75, 76] but have until recently only attracted little attention in fungi and hence only few fungal examples are characterized.

The detoxification strategy is essentially a mechanism where the fungus possesses an additional enzyme modifying the chemical structure of the SM thereby preventing the molecule from binding to its target (or making the binding less effective). An example of this strategy is found in the gliotoxin cluster in *A. fumigatus*, where GliT (an oxidoreductase) modifies the gliotoxin structure creating a less toxic compound within the producing fungi whereas other fungi do not have this mechanism and are therefore attacked by the toxin [77, 78].

The transportation strategy covers both exportation and compartmentalization of the compound. This is one of the best known strategies fungi use to get rid of toxic materials in general and this mechanism is also found in biosynthetic gene clusters. Transporter genes such as the ABC superfamily and major facilitator superfamily (MFS) are often found in biosynthetic gene cluster. These can have many functions such as transportation of precursors into the cells, transportation of intermediates to subcellular compartments, ensuring that the compounds reach the extra-cellular environment or as self-protection mechanism transporting the toxic compound out of the cell or into a specific compartment [63]. In the gliotoxin cluster, the transporter gliA encodes a highly effective efflux pump which is involved in resistance to gliotoxin [78]. *Tri12* encodes a MSF-type transporter located within the trichothecenes cluster and it has been shown to be involved in self-resistance [79, 80]. There are also several examples of transporters not located within a biosynthetic gene cluster conferring resistance [81]. The localization within the cell is a less studied aspect of secondary metabolite biosynthesis, but recently it has been shown to be an important factor in successful SM biosynthesis and in self-resistance [63, 82]. One example is aflatoxin production in *A. parasiticus*, which is carried out in several different vesicles including the specialized aflatoxisomes thought to confer self-resistance [83, 84, 63]. The production of penicillin, deoxynivalenol and trichothecene have also been shown to be compartmentalized [82, 63].

The third self-resistance mechanism is 'duplication of the target gene'. The resistance can originate from two different mechanisms; sheer numbers – having overexpressed and hence more of the target protein or insensitivity – or having a modified insensitive version of the target gene. There are several examples of

clusters containing a second version of the target genes including the lovastatin, compactin, and fumagillin clusters [85, 71, 86]. These are hypothesized to confer resistance, but whether the mechanism is due to more copies or if the duplicated version is insensitive is not known. The mycophenolic acid cluster was identified based on the resistance gene, a duplication of the target gene, IMP dehydrogenase (IMPDH) [87]. It has later been shown that the second copy is an insensitive version conferring resistance [88, 89]. More recent examples of this mode of resistance include the fellutamide B cluster [90] and the aspterric acid [91] indicating that this mode of resistance might be more widely distributed than previously known.

Despite utilizing one of these strategies the producing fungi can also be slightly affected since the resistance mechanism might ensure survival but not completely unaffected cells. However as long as the producing fungi is surviving better than the others it is a viable tactic.

2.4 Manuscript I – Review: linking compounds and gene clusters

Only a small fraction of the enormous natural resource of biological and chemical diversity have been explored thus leaving a large untapped potential. In order to exploit this resource for novel pharmaceutical and industrial uses and creating economically viable cell factories it is important to establish the link between the compound and the gene clusters. Here we present a review, Manuscript I, where we have outlined various strategies for linking biosynthetic gene cluster to compounds and vice versa.

Manuscript I will be submitted to Fungal Genetics and Biology.

Strategies to establish the link between biosynthetic gene clusters and compounds

Inge Kjæbølling[a], Uffe H. Mortensen[a], Tammi Vesth[a], Mikael R. Andersen[a]

[a] *Technical University of Denmark*

Abstract

Filamentous fungi produce a vast number of bioactive secondary metabolites (SMs) where some have found applications in the pharmaceutical industry including antibiotics and immunosuppressants. As more and more species are whole genome sequenced the number of predicted clusters is ever increasing - holding a promise of novel useful bioactive SMs. To be able to fully utilize the potential of novel SMs it is necessary to link the SM and the genes responsible for producing it. This can be very challenging but there are many strategies and tools developed for this purpose. In this review we provide an overview of the methods used to establish the link between SM and biosynthetic gene cluster and vice versa, along with the challenges and advantages of each of the methods.

Keywords: Secondary Metabolites, Biosynthetic Gene Clusters, Filamentous fungi

1. Introduction

Fungal secondary metabolites (SMs) are a unique and extraordinary source of bioactive SMs including both medically utilized and novel promising SMs. The best known exploited fungal bioactive SMs includes antibiotics such as penicillin and cephalosporin, hypercholesterolaemic agents such as lovastatin, immunosuppressants such as cyclosporinand mycophenolic acid, as well as antifungals such as echinocandin and derivatives [54, 47, 86]. Besides the SMs with beneficial properties, fungal SMs also includes harmful toxins. The most carcinogenic toxins aflatoxins are produced by members of the genus *Aspergillus* [10]. Other toxins are detrimental to human, animal and plant health including fumonisins, ochratoxin and gliotoxin [10, 59].

The genes responsible for producing the SMs are arranged in clusters on the genome usually containing an enzyme creating the core structure of the compound (backbone enzymes), tailoring enzymes, and potentially regulatory enzymes and resistance genes, such as transporters [79, 20]. The most common backbone enzymes are polyketide synthases (PKSs) or a non-ribosomal peptide synthetases (NRPSs), which both have highly conserved domains.

Email addresses: ingek@bio.dtu.dk (Inge Kjæbølling), um@bio.dtu.dk (Uffe H. Mortensen), tcve@bio.dtu.dk (Tammi Vesth), mr@bio.dtu.dk (Mikael R. Andersen)

Due to the highly conserved domains of the backbone genes and the arrangement of the genes in clusters it is possible to predict secondary metabolite gene clusters in genomes with sequence domain and knowledge-based algorithms such as SMURF and antiSMASH [55, 15]. There are however also secondary metabolite gene clusters (SMGCs) that do not follow the conventions and therefore they are difficult to predict using these algorithms, this includes tryptoquivaline in *A. clavatus* [43] and the echinocandin in *Emericella rugulosa* [23] which are split in two parts. Some cluster types are not easily detected by the prediction algorithms such as terpenoid based clusters since the terpene synthase are not as conserved as PKS and NRPS. Extra genome mining have to be done to identify them as demonstrated by Bromann et al. [21].

With an ever increasing number of whole genome sequenced fungal strains and species, it is clear that the number of SMGCs far exceeds our expectations from known SMs and the number of expressed clusters. This indicates a big potential of novel useful bio-active SMs.

To fully understand and exploit the rich resource of fungal SMs, it is important to establish the link between the SM and the SMGC, both to be able to discover novel SMs but also to optimize production of the SMs, to make it feasible for an industrial setting. Knowing which genes are responsible for the production of a SM makes it possible to use metabolic engineering strategies to make an economically viable process which is essential in unlocking the potential of fungal SMs. Here we present an overview of various strategies that can be used to discover the connection between a SM and gene cluster. The aim of this review is to give an overview of strategies that have been used to study secondary metabolism in fungi, illustrated by selected examples.

This review consists of two main parts: 1) Forward strategies going from cluster to SM and 2) Reverse strategies going from SM to cluster. The forward strategies are divided into two parts: native and heterologous strategies. These are further divided into strategies for active and silent clusters for native hosts, while the heterologous strategies are divided into hosts and constructs. The reverse strategies are divided into three parts depending on the approach: 1) Homology search 2) Retro-biosynthesis and 3) Comparative genomics.

2. Forwards strategies from cluster to secondary metabolite

In our definition, the starting point of the *forward strategies* is an identified biosynthetic gene cluster of interest and a wish to elucidate the produced secondary metabolite (SM). At this initial point the first thing to consider is the species the cluster is identified in. This is crucial for which strategy to use – native or heterologous – expression. Each of these strategies will be outlined and discussed in detail below.

2.1. Strategies in native hosts

If the secondary metabolite gene cluster (SMGC) of interest is found in a suitable host, with established molecular tools and is cultivatable in the laboratory, there are several strategies that can be employed. Which to choose depends mainly on the native activity of the cluster, if the cluster is expressed it is called active and if it is not expressed it is denoted silent. This information is often not known initially, but can be investigated directly by

reverse transcription polymerase chain reaction (RT-PCR) or globally with transcriptomics or proteomics. The examination of a SMGC is usually not a single process and often it requires a combination of various strategies to clarify the link between secondary metabolite and secondary metabolite gene cluster and to elucidate the biosynthesis.

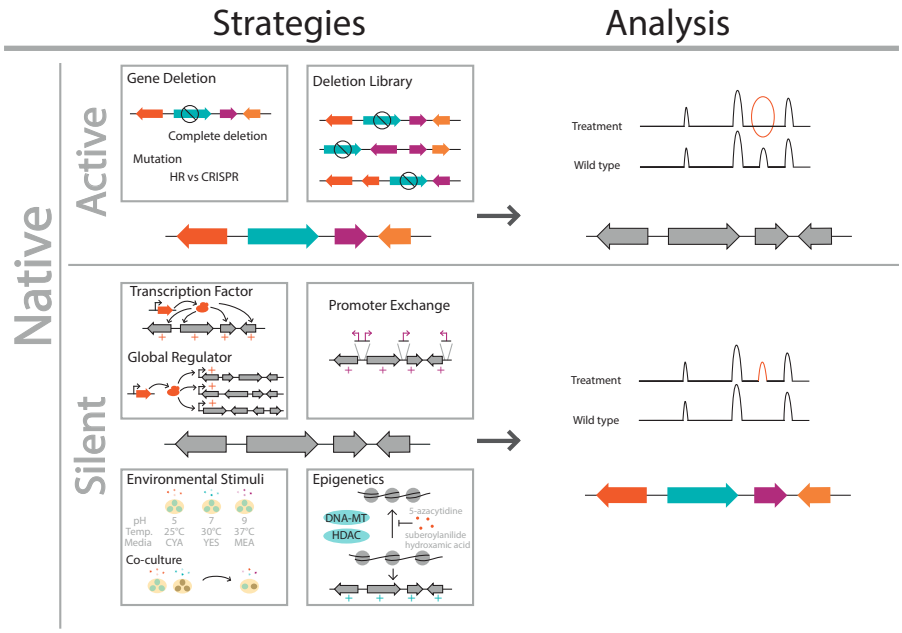


Figure 1: Strategies for investigating a cluster in the native host. On top are the strategies useful for active (expressed) secondary metabolite gene clusters, A1 – gene deletion and A2 – the generation of a gene deletion library. At the bottom the strategies used for silent (not expressed) gene clusters are illustrated, S1 – overexpression of a global or cluster specific transcription factor, S2 – promoter exchange of all the biosynthetic genes, S3 – using various environmental stimuli (change pH, tempearure, media components or co-culture with other microorganisms) to change the expression profile, S4 – epigenetics or chromatin remodelling wither by adding compounds monipulating the epigenome or by gene deletions of epigenetic regulators.

2.1.1. Gene deletion or disruption strategies in active clusters

Cluster specific gene deletions or disruptions. If a secondary metabolite gene cluster (SMGC) is active under some known condition, it is possible to make gene deletions, disruptions or silencing in the cluster to elucidate the pathway or simply to couple the cluster to a secondary metabolite (SM), see Figure 1 A1. By comparing the chemical spectrum of a wild type with the modified strain, it is possible to identify the SM missing in the modified strain and potentially identify intermediates in the biosynthetic pathway.

The strategy of gene deletions (Figure 1 A1) has been used in countless studies. Disruption of the *FUM5* gene combined with complementation studies revealed the involvement in fumonisin production in *Gibberella fujikuroi* [84]. Later *FUM6-FUM9* were identified in *Fusarium verticillioides*, analysed and disrupted showing that these genes are also involved in Fumonisins production [93]. To investigate the red pigment aurofusarin produced by *Fusarium pseudograminearum* and *F. graminearum* Malz et al. created aurofusarin deficient mutants using first random mutagenesis revealing a locus of interest including a PKS, *PKS12*. To confirm the *PKS12* gene involvement in aurofusarin production a targeted gene disruption was performed [66]. The cyclopiazonic acid (CPA) biosynthetic genes were identified in *Aspergillus flavus* where three genes were disrupted and two of these mutants completely abolished CPA production [25]. In *A. niger*, the *albA* gene was shown to be responsible for the production of pigments as well as naphtho-γ-pyrones through a deletion study. This helped identify the link between SM and genes. In addition, it created a useful strain for further analysis since some of the main SMs produced by *A. niger* are disposed of, resulting in a more clean background strain [26]. In some cases gene deletion or disruption has turned out not to be possible and RNA silencing has been used as an alternative. This was seen in a study identifying the cytochalasin gene cluster in *Penicillium expansum* [92].

The advantage of the gene deletion strategy is that it is a straight-forward approach often resulting in establishing links between secondary metabolites and the responsible genes, while the drawbacks are that it requires an expressed cluster, a cultivable organism and molecular tools. Many clusters are not active under standard laboratory conditions and several studies have thus combined deletion strategies with other methods such as transcription factor overexpression. One example of this is a study by Neubauer et al. where a transcription factor was overexpressed to activate a cluster, followed by gene deletions thereby identifying the ergochrome gene cluster in *Claviceps purpurea* [74]. More examples will be given under the respective combination strategy in the following sections. Another limitation of the deletion strategy is that it is highly sensitive to the methods used both during extraction and isolation. The SM of interest needs to be detectable with the extraction method used and isolated in amounts above the detection limit of the instrument. It is therefore essential to select which methods to use depending on the SM of interest.

Gene deletion libraries. A further sophistication of the deletion strategy focusing on one secondary metabolite gene cluster (SMGC) of interest is the generation of deletion libraries, Figure 1 A2. Nielsen et al. generated a polyketide synthases (PKS) deletion library, deleting each of the 32 predicted PKSs in *A. nidulans* where a few were already known. By growing the mutants on various media and comparing with the reference, they were able to identify PKSs involved in arugosins, violaceol, austinol and dehydroaustinol biosynthesis [75]. In a similar study in *Gibberella zeae*, 15 PKSs were disrupted and known SMs such as zearalenone, aurofusarin, fusarin C and the black perithecial pigment were linked to specific genes [41]. Another deletion library focused on the regulatory genes in *A. nidulans*, where 128 kinases were deleted. In this study it was observed that the secondary metabolism was affected in several of the mutants [33]. Yaegashi et al. screened this knockout library for changes in secondary metabolism. They found the SM aspernidine A and identified the secondary

metabolite genes in *A. nidulans* [109].

Using a panel of deletion mutants in the manner presented above has a great potential for future studies, given the developments in genome editing technologies for fungi sparked by the CRISPR-Cas9 technologies [78].

2.1.2. Strategies for triggering activation of silent secondary metabolite gene clusters

With the sequencing of the first fungal genomes, one of the major discoveries was that the number of secondary metabolite gene clusters (SMGCs) greatly outnumbered the known SMs thereby revealing an even bigger resource but also showing that most SMGCs are not active or active enough to produce a detectable amount of SM during standard laboratory cultivation [54, 77, 80, 91]. The gene clusters that are not expressed during standard conditions are often referred to as silent, while gene clusters where the product is unknown is referred to as orphan or cryptic clusters. With a large number of predicted but unexplored clusters, efforts have gone into the development of strategies for activating these silent gene clusters. The strategies include global changes such as global regulators, modified growth conditions and epigenetics plus cluster specific strategies such as promoter replacement and transcription factor overexpression.

Global regulators of Secondary Metabolism. Firstly we will focus on global regulation, Figure 1 S1. Bok et al. identified the protein LaeA regulating secondary metabolite production in several *Aspergillus* species [18]. In *laeA* deletion strains, the production of sterigmatocystin and penicillin and gliotoxin was decreased in *A. nidulans* and *A. fumigatus* respectively, whereas in *A. terreus* overexpression of *laeA* increased the production of lovastatin. LaeA was thus established as a global regulator of secondary metabolism in *Aspergillus* species [18]. In a following study, *laeA* deletion and overexpression strains were used to identify active clusters through expression analysis in *A. nidulans*. A gene deletion in one of the clusters revealed that the cluster is responsible for producing terrequinone A [17]. The study of LaeA has been expanded to many other filamentous Ascomycetes where the link between LaeA and secondary metabolite production has been studied for instance in the study of T-toxin in *Cochliobolus heterostrophus* [14] and bikaverin, fumonisins, fusaric acid and fusarins in *Fusarium verticillioides* [22] just to mention a few. For an excellent review of this topic please refer to Jain and Keller [51].

It has been shown that the LaeA is part of the velvet complex with VeA and VelB which connects light-response, developmental regulation and regulation of secondary metabolism [8]. In a study comparing the transcriptional profile of an *A. fumigtus* wild type, *laeA* deletion and a complementation control strain showed that 13 out of 22 SMGCs were positively regulated by LaeA. Of the LaeA-regulated clusters 54% are located within 300 kb of telomeres [81]. This could suggest a relationship between LaeA activity and chromatin modification which has been hypothezised by Keller et al. [54] however the hypothesis that LaeA methylates histones has not yet been verified nor refuted [51].

Several other proteins have been shown to have general regulatory functions affecting secondary metabolism these includes the nitrogen regulator AreA [52, 70, 103], the pH regulator PacC [37] and the carbon regulator CreA [35, 36]

The immediate advantage of using global regulators is that more SMGCs are expressed and more metabolites are produced, the main disadvantage is that the effect is not specific, thus many clusters are affected which can make the chemical analysis difficult. If a specific cluster of interest is investigated it is not certain that it will be affected by the global regulator.

Cluster specific transcription factors. A more targeted strategy connected to regulation is the overexpression of a cluster-specific transcription factor (TF), Figure 1 S1. The method was first applied by Bergmann et al. in *A. nidulans* where the cluster specific TF was integrated ectopicaly under the control of an inducible promoter. This led to the elucidation of the novel SMs aspyridones A and B and identification of the PKS-NRPS hybrid SMGC responsible for production [12]. Subsequently, several SMs and clusters have been linked and characterized using a similar strategy e.g Asperfuranone in *A. nidulans* [28], a diterpene in *A. nidulans* [21] and Azaphilone in *A. niger* [112]. In *Fusarium fujikuroi*, a pathway specific TF and the PKS backbone enzyme was overexpressed resulting in the production of 4 new SMs, fujikurins A-D [52, 98]. Scopularide A and the responsible cluster was identified and verified by overexpression of the cluster specific transcription factor in a marine-derived *Scopulariopsis brevicaulis* strain [63].

It is clear that the overexpression of a cluster-specific regulator can be a powerful strategy. However, it is not always straightforward. In a cluster containing two NRPS genes and a regulatory gene named *scpR* (for secondary metabolism cross-pathway regulator), the overexpression of *scpR* led not to the activation of the NRPSs as expected, but to the activation of the asperfuranone cluster located on another chromosome in *A. nidulans* [11]. This showed that cluster-specific TFs are not necessarily located within the cluster it is affecting and thus adds another layer of complexity to SMGC regulation.

The main advantage of the method is that it – when successful – allows coordinated expression of all SM pathway genes. Other advantages of expressing a cluster-specific TF include that only one gene has to be manipulated and the integration can be ectopic which evades the limitations of homologous recombination. The main limitation of this method is that it requires the cluster of interest to have a pathway specific TF.

Promoter exchange. When no cluster specific transcription factor is available, another option is to replace the promoters of all the genes in the SMGC to force the expression of the cluster genes, Figure 1 S2. This was done by Yeah et al. in 2016 for the fellutamide B cluster in *A. nidulans*, where the promoter of six genes were replaced with the regulatable *alcA* promoter by recycling a selectable marker [110]. A similar strategy was used in *A. nidulans* for the cluster responsible for a conidiophore pigment where the promoters of three genes (*ivoA-C*) were replaced with the *alcA* promoter [97].

The main advantage of this method is that it is specific and can be used for all clusters, no transcription factor is required. Disadvantages are that it requires molecular tools in the organism and it is quite labour intensive especially if the cluster consists of many genes.

Epigenetics / chromatin remodelling. As mentioned in connection with LaeA, chromatin remodelling has been shown to be involved in regulation of secondary metabolites (SMs). The

relationship between SM and chromatin structure can be exploited to activate otherwise silent gene clusters. The epigenetic landscape can both be modified by molecular methods (e.g. gene deletion) or chemically by adding small molecules manipulating the fungal epigenome.

In 2007, Shwab et al. deleted *hdaA* encoding a histone deacetylase (HDAC) in *A. nidulans* showing that the levels of the two sub-telomeric cluster products, penicillin and sterigmatocystin, increased while terraquinone A levels were unaffected and this cluster is not subtelomeric. To investigate the mechanism in other fungi, *A. alternate* and *P. expansum* were treated with a HDAC inhibitor (Trichostatin A) which resulted in significant increased levels of several SMs in both species. This study showed that HdaA plays an important role in suppression of SM located in the sub-telomeric regions and that it might be a conserved mechanism across the fungal species [94]. In a study by Williams et al. it was shown that small-molecule epigenetic modifiers were effective in eleven out of 12 species tested where one or more modifiers caused the production of new SMs or enhanced production of known SMs compared to untreated controls. Two species were investigated further, *Cladosporium cladosporioides* and *Diatrype disciformis* and several new SMs were identified and characterised in each species showing the applicability of the method [104]. In a study of *A. niger* it was shown that many of the clusters were affected by deacetylase inhibitors [39] illustrating the potential of chromatin remodelling in the activation of silent gene clusters. Again it was also shown that the affect was mainly on clusters located near the telomeric regions.

CclA is a part of the complex COMPASS, an eukaryotic transcriptional effector methylating lysine 4 of histone H3 (H3K4)1 thus affecting chromatin-mediated processes. The deletion of CclA, in *A. nidulans* showed activated expression of otherwise silent SMGC including the new cluster generating monodictyphenone, emodin and emodin derivatives, plus a cluster encoding two anti-osteoporosis polyketides, F9775A and F9775B [16].

Many of the advantages and disadvantages of this strategy are similar to the global regulator strategy, however comparing those two strategies, the advantage here is that it can be made in species without molecular tools if using small molecule inhibitors.

Environmental stimuli. Secondary metabolite gene clusters (SMGCs) are induced at specific conditions, under which conditions they are active can be identified by changing the environment or adding various stimuli. Gressler et al. had tried expression of a transcriptional activator in the cluster of interest in *A. terreus*, but without success, instead they turned to environmental stimuli. A combination of metabolic profiling, monitoring gene expression, and a *lacZ* reporter strain was used to identify on which media the cluster was active and this strategy finally led to the identification of isoflavipucine and dihydroisoflavipucine [45].

It has further been shown that ionic liquids can stimulate the secondary metabolism in *A. nidulans* where 32% of described backbone genes were upregulated whereof several were normally silent [4].

Co-cultivation of fungal species with bacteria is another environmental stimulus that has been used successfully to activate silent gene clusters. Wakefield et al. grew *A. fumigatus* MR2012 with isolates of *Streptomyces leeuwenhoekii* which led to the production and identification of luteoride D (a luteoride derivative) and pseurotin G (a pseurotin deriva-

tive) plus production of SMs not previously identified in this species (terezine D and 11-O-methylpseurotin A) [99].

The advantages of using environmental stimuli is similar to global regulator and chromatin remodelling in that more clusters are activated however again it is not specific and it is not known if the cluster of interest will be affected. Another advantage is that it is possible to use this strategy for unexplored species, since no molecular tools are needed.

Future possibilities in native strategies. With the advancement of molecular tools such as CRISPR-Cas9 (Clustered Regularly Interspaced Short Palindromic Repeats) in fungi [78], many of the native strategies will be applicable in more non model species. Another future possibility based on the CRISPR-Cas system is the development of synthetic transcription factors where a deficient Cas9 can be fused to an effector domain thus inducing transcription of specific genes which has already been applied in yeast [38].

2.2. Heterologous Strategies

Heterologous expression strategies can be applied if the cluster of interest is found in a species not easily genetically manipulated, or if the native strategies have been tested and failed. The first thing to consider with heterologous expression is which host to select. This choice depends on many factors. We will here focus on three categories of hosts; bacterial, yeast and filamentous fungi. The second thing to consider is the construct and how comprehensive it should be, including only backbone synthase, promoter, swap of transcription factor or engineering of the whole cluster with new promoters etc. for all genes in the cluster. Which construct to use depends on the cluster and the aim. In the follwing section the three categories of hosts are examined follwed by a section covering the construct

Bacterial host. The first group of hosts we will consider is bacterial hosts such as *Escherichia coli* and *Streptomyces coelicolor*. *S. coelicolor* has been used as a hosts to express the 6-methylsalicylic acid synthase gene from *Penicillium patulum* [9]. Later the 6-methylsalicylic acid synthase gene from *P. patulum* was successfully expressed in *E. coli* and *S. cerevisiae* producing 6-methylsalicylic acid [53].

In a more recent study the gene encoding the BbBEAS nonribosomal peptide synthetase isolated from *B. bassiana* was heterologously expressed in *E. coli* which was then able to produce beauvericin when the precursor D-Hiv was added [48, 108]

Several of the tropolone biosynthetic genes from *Talaromyces stipitatus* were expressed in *E. coli* and purified for in vitro analysis thus verifying the expected biosynthetic pathway [32]. The PKS4 protein from *Gibberella fujikuroi* was expressed in *E. coli*, purified and in vitro analysis was performed showing a functional enzyme producing the secondary metabolite SMA76a [65]. Similarly the LovD protein from lovastatin biosynthetic cluster was expressed in *E. coli*, purified and analysed in vitro [105]. In an investigation of the basidiomycete *Coprinus cinereus*, six sesquiterpene synthases were expressed in *E. coli* and 5 of the products were identified directly from *E. coli* cultures using GC-MS [2]. Using the same approach 11 sesquiterpene synthetase from *Omphalotus olearus* were characterised by heterologous expression in *E.coli* [102].In the investigation of the Azaphilone gene cluster the

gene *AzaH* (a FAD-dependent monooxygenase) was expressed in *E. coli* in order to make in vitro assays of the protein and investigate its function in the azaphilone biosynthesis [112].

E. coli has several advantages as a host for heterologous expression; 1) easy to culture and fast growing 2) a well developed molecular toolbox 3) a well-understood primary metabolism and 4) the absence of endogenous secondary metabolite pathways thus limiting the risk of cross-talk and interference with native proteins [44, 82]. There are however also several challenges when heterologously expressing fungal secondary metabolite genes in bacterial hosts. The challenges include; 1) bacteria's inability to process eukaryotic introns which thus have to be eliminated 2) codon bias can cause problems in expression 3) correctly folding of the synthesized proteins 4) required post-translational phosphopantetheinylation/modifications and 5) availability of the precursors [3, 44, 48]. Bacterial hosts are therefore most often used for *in vitro* enzyme analysis of a specific biosynthetic gene.

Yeast as a heterologous host. The second host we will consider is yeast, in particular *Saccharomyces cerevisiae*, which taxonomically is closer to filamentous fungi than bacteria, belonging to the same kingdom.

As mentioned earlier the gene encoding 6-methylsalicylic acid synthase from *P. patulum* was expressed in a *S. cerevisiae* strain including a heterologous phosphopantetheinyl transferase which creates the active holo PKS from the apo-PKS [60, 53]. From this strain 6-methylsalicylic acid was produced and the amount was twice as high as in the native species and much higher than in *E. coli* [53].

The lovastatin nonaketide synthase, LovB from *A. terreus* was expressed in an engineered strain of *S. cerevisiae* containing the a phosphopantetheinyl (ppant) transferase gene *npgA* from *A. nidulans* [71], in order to perform in-depth *in vitro* investigation of the catalytic function and mechanism [64].

Ishiuchi et al. engineered a *S. cerevisiae* strain to include *matB* (malonyl-CoA synthetase) and *npgA* (phosphopantetheinyl transferase) and successfully used this for expression of five PKS and one NRPS and characterization of the produced SMs [49]. Several other studies have used optimized versions of *S. cerevisiae* for heterologous expression of synthase genes, for identification of the products and characterization of the mechanisms, including identification of 10,11-Dehydro-curvularin and characterization of a mechanism for aryl-aldehyde[100, 107]

Not only synthase genes have been expressed in *S. cerevisiae*, but also whole clusters. The biosynthetic genes from the hypothemycin gene cluster from *Hypomyces subiculosus* were for instance expressed in a PKS optimized yeast strain [53, 85] and based on these experiments it was possible to propose a biosynthetic pathway [85]. In another study by Rugbjerg et al. three biosynthetic genes from *Fusarium graminearum* were co-expressed with the phosphopantetheinyl transferase (npgA) gene from *A. fumigatus* resulting in the production of Rubrofusarin [89].

In a recent study, Harvey et al. developed HEx (Heterologous EXpression) synthetic biology platform for fast and scalable expression of fungal biosynthetic genes and their encoded metabolites in *S. cerevisiae* [46]. In this study 41 biosynthetic gene clusters from diverse ascomycete and basidiomycete fungal species were expressed in *S. cerevisiae* and

54% resulted in SMs not natively found in yeast [46]. This platform brings the method of heterologous expression in yeast a big step forward and potentially opens the door to discovery of many natural products.

Using *S. cerevisiae* as a heterologous host has many advantages, some of which are similar to those mentioned for *E.coli*, but there are additional advantages to using *S. cerevisiae*: 1) *S. cerevisiae* is a unicellular organism, easy to culture and it grows faster than most filamentous fungi; 2) Powerful genetic tools have been developed for protein expression and pathway construction, including highly efficient homologous recombination; 3) Native secondary metabolism is very limited in *S. cerevisiae* thus minimizing the background and potential cross-talk [95]; 4) The building blocks for polyketide synthesis such as acetyl-CoA and malonyl-CoA plus cofactors such as NADPH and S-adenosylmethionine are naturally produced in yeast; 5) Lastly yeast also belongs to the fungal kingdom and it can typically produce tailoring enzymes and support correct folding [19, 3] .

The challenges of using *S. cerevisiae* as a heterologous host includes: 1) A heterologous gene is required for activation of the synthase such as the surfactin P-pant transferase (Sfp) from *Bacillus subtilis* [53]; 2) *S. cerevisiae* has different and few introns which can cause problems in mRNA splicing [58]; 3) In yeast codon usage is biased towards AT which can cause problems if the gene of interest is GC rich [73]; A low or even lacking production of required precursors and building blocks [53, 73]; 4) Lack of compartmentalization which might be important for SM production [88]; 5) Lastly there is a risk of toxicity of the produced SM.

Filamentous fungi as heterologous host. Many of the challenges seen for yeast and bacterial hosts can be overcome by using a filamentous fungal hosts. The model fungus *Aspergillus nidulans* is often used, because it has a well developed genetic toolbox, also *A. oryzae* which has a limited endogenous secondary metabolism is often employed.

Initially many of the studies of heterologous expression were based on a single gene, often the backbone enzyme. One example is the study of *albA* from *A. fumigatus* which is involved in conidial pigment biosynthesis and it was heterologously expressed in *A. oryzae* to show that the PKS is a naphthopyrone synthase (expected based on the sequence similarity) and not a tetrahydroxynaphtalene synthase (expected from the color) [101]. The drawback of single gene studies is that most often they do not give the final secondary metabolite (SM) or the biosynthetic pathway.

Other studies have expressed whole clusters in heterologous hosts, Smith et al. was one of the first to do this. They cloned the penicillin biosynthetic gene cluster from *Penicillium chrysogenum* on to a cosmid vector, transferred it to *Neurospora crassa* and *A. niger* and showed that penicillin was produced [96]. This approach however relies on the native promoters functioning in the new host and correct mRNA processing.

One method of circumventing this dependency is to put the transcription factor under a strong promoter. This approach was seen in a study of a cryptic polyketide cluster from *Trichophyton tonsurans* where four biosynthetic genes of the cluster were expressed from their own promoters, and only the cluster specific transcription factor promoter was replaced with the strong *A. nidulans* gpdA promoter [111]. A neat detail in the design of this study

was that the cluster was inserted in the *wA* locus of *A. nidulans* encoding a pigment PKS, facilitating the screening of correct recombination.

Many studies have been conducted investigating biosynthetic gene clusters by heterologous expression including pyripyropene from *A. fumigatus* expressed in *A. oryzae*[50], the citrinin cluster from *Monascus purpureus* expressed in *A. oryzae* [90] and the *Pfma* cluster from *Pestalotiopsis fici* expressed in *A. nidulans* synthesizeing the melanin 8-dihydroxynaphthalene (DHN) [113] just to mention a few.

Heterologous expression in filamentous fungi can give rise to some challenges, the host can be affected by the inserted cluster causing a changed metabolite profile or cross-talk between the inserted cluster and native clusters can arise which makes it difficult to identify the correct new SM. This was illustrated by a study expressing a polyketide gene cluster originating from a fungal endophyte in *Fusarium verticillioides* [106], where the main product identified was fusaric acid, which is a mycotoxin normally found in *Fusarium* species.

There are several studies optimizing heterologous hosts to make the strategy more effective. Chiang et al. developed an optimized heterologous expression system in *A. nidulans* [27]. Their first step was to delete native biosynthetic gene clusters (of sterigmatocystin, emericellamide, orsellinic acid/F9775A,B, asperfuranone, monodictyphenone, and terrequinone) in order to reduce the SM background and facilitate detection of novel products plus to increase the pool of pre-cursors for the desired products. Next they developed a method for heterologous expression of biosynthetic genes using a system of a recycable marker thus permitting the expression of entire clusters which was shown for the asperfuranone cluster from *A. terreus*.

A more recent effort by Clevenger et al. using fungal artificial chromosomes (FACs) and metabolomic scoring (MS) have made the it possible to scale up the analysis which was demonstrated by investigating 56 SMGCs originating from *A. terreus, A. aculeatus* and *A. wentii* in *A. nidulans* [30]. In the study, they detected 17 SMs produced by 15 different FACs. In a subsequent study the FAC-MS method was used to elucidate the biosynthesis of acu-dioxomorpholine [87].

The advantages of using a filamentous fungi as host includes; 1) The genetic systems are generally compatible correctly translation folding and post-translational modifying the foreign gene(s) hence obviating the need for codon optimization, intron removal etc. 2) The secondary metabolite machinery is present, making most common precursors available. The downside is that there is a lot of background which can make the chemical analysis difficult and cause cross-chemistry making it complicated to identify the SM produced by the inserted genes. An additional disadvantage is that it is very time consuming using filamentous fungi as hosts since they are challenging to engineer and slow growing.

Design of constructs. Besides selecting the host for heterologous expression, another ting to consider is the construct as mentioned earlier. Expressing a gene heterologously either depends on the host having similar promoters and terminators and mRNA transcript processing or it requires extensive engineering of the cluster. The design of the constructs depends on the cluster of interest, the host and the aim. Here we have divided it into three main strategies: 1) Expression of the synthase/synthetase, 2) Inserting the natural cluster,

potentially with engineering of the transcription factor and 3) Engineering of the whole cluster including new promoter etc. for all genes.

In some initial studies of cryptic biosynthetic gene clusters only the synthase has been expressed. This method is used to investigate the core structure of the secondary metabolite (SM) to give an initial indication of what structure is produced and can also be used in screening studies. Expressing only the synthase has the advantage that it is most often only one gene, making it easier and less labour intensive to exchange promoter and occasionally terminator. This strategy was used in the study of six sesquiterpene synthases from mushroom-forming fungi (*Agaricomycetes*) where the sesquiterpene synthases were expressed in *E. coli* and/or *S. cerevisiae* thereby characterizing the enzymes and identifying the major sesquiterpene hydrocarbons produced [2]. Ishiuchi et al. expressed five PKSs and an NRPS in *S. cerevisiae* and identified the corresponding natural products [49]. Likewise two polyketide synthases responsible for cladosporin production was expressed in *S. cerevisiae* to confirm the involvement in cladosporin production and understand the mechanism and biosynthesis [31]. Munawar et al. expressed a nonribosomal peptide synthetase from *Fusarium sacchari* in *A. oryzae* and showed that it is responsible for producing the sideriohore ferrirhodin [72]. The disadvantage is that it does not give the final SM but it gives the core structure of the SM and the initial step in the biosynthesis.

most often several genes have to be expressed in order to analyze an entire biosynthetic cluster. Examples have been seen where an entire gene cluster is transferred without modifying it, this requires that the heterologous host is closely related to the native species so that the transcriptional and translational machinery is compatible, however the amount of produced SM is often quite low. The citrinin gene cluster from *Monascus purpureus* was expressed in *A. oryzae* without modifying the cluster resulting in low production [90]. If the cluster contains a cluster specific transcription factor (TF) it is possible to only exchange the promoter of the TF and get increased expression of the rest of the cluster, similar to the strategy of activating silent gene clusters in section 2.1.2. This strategy was employed on the citrinin gene cluster which contains a TF, this way an almost 400-fold higher citrinin production was achieved [90]. Similarly, the geodin cluster from *A. terreus* was successfully expressed in *A. nidulans* by replacing the promoter of the TF with a strong constitutive promoter [76].

The advantage of this strategy is that it produces the final SM of the cluster and it requires minimal engineering of the cluster. The disadvantage is that it requires compatible cellular machinery of the native and heterologous host or the presence of a cluster specific TC.

If the cluster of interest does not contain a TF, the third approach can be utilized which is engineering of the entire cluster exchanging all the promoters. Four biosynthetic genes from *Phoma betae* was expressed in *A. oryzae* in a vector-based approach using the starch-inducible promoter/terminator from the *amyB* gene thereby producing aphidicolin [40]. Bailey et al. successfully expressed seven biosynthetic genes from the basideomycete *Clitopilus passeckerianus* in *A. oryzae* using constitutive *A. oryzae* promoters thereby producing pleuromutilin [6].

The advantage of engineering the entire cluster is that there is no requirements or re-

strictions as to which clusters can be investigated. The drawback is that it is a more labour intensive method. There are several excellent reviews on the subject of heterologous expression, for further reading please refer to [3, 5, 61].

Future perspectives. As the price of synthesis of long stretches of DNA keeps dropping due to technical advancement, novel strategies are emerging, making it possible to circumvent several time consuming cloning steps and thereby easier to investigate selected clusters or screen more clusters. Especially the strategy of promoter swapping of all genes in a cluster becomes much more feasible when using synthetic DNA.

3. Reverse strategies going from secondary metabolite to cluster

An alternative to the approach of starting from an interesting cluster and linking it to a secondary metabolite (SM), is to start from a SM of interest. If a SM has been identified in a certain species, there are several methods for identifying the responsible biosynthetic gene cluster. Here we have divided it into three main approaches 1) Homology search 2) Retro biosynthesis and 3) Comparative genomics (Figure 2). Which strategy to use essentially depends on the initial knowledge. As seen in the forward strategies, a combination of several strategies is often needed to identify and verify the secondary metabolite gene cluster (SMGC). All the strategies mentioned here requires whole genome sequences of the producing organism. These methods has thus only been made possible in the post-genomic era. As whole genome sequences are becoming more and more attainable, the use of these strategies will surely only increase.

3.1. Homology Search

The first strategy presented here is based on homology search, Figure 2 strategy 1. The starting point is a known secondary metabolite (SM) where the same or a similar SM is produced in another species where the secondary metaboolite gene cluster is known. In this case, the known cluster genes can be used to search for similar genes in the selected organism. The most important is the synthase/synthetase responsible for producing the backbone of the SM. A derivative of this strategy (genetic dereplication) has a wider scope, not focusing on a single SM but instead identifying homologs of all characterized biosynthetic gene clusters. When a novel species have been whole genome sequenced this approach is very useful to identify the known and novel clusters.

The ochratoxin gene cluster was identified in *A. carbonarius* using homology search of the ochratoxin cluster predicted in *A. niger*. The following deletion of a PKS in the predicted cluster eliminated all production of ochratoxin confirming the biosynthetic role[42]. The method of homology search can also be used to find the putative clusters for similar SMs that could use parts of the same biosynthetic pathway. This was shown for in the identification of the novofumigatonin cluster in *A. novofumigatus* which was identified based on homology to another meroterpenoid, the terretonin cluster from *A. terreus* [57, 67].

Using homology search, it is also possible to investigate the secondary metabolite potential of a newly sequenced species. The secondary metabolism was investigated in the

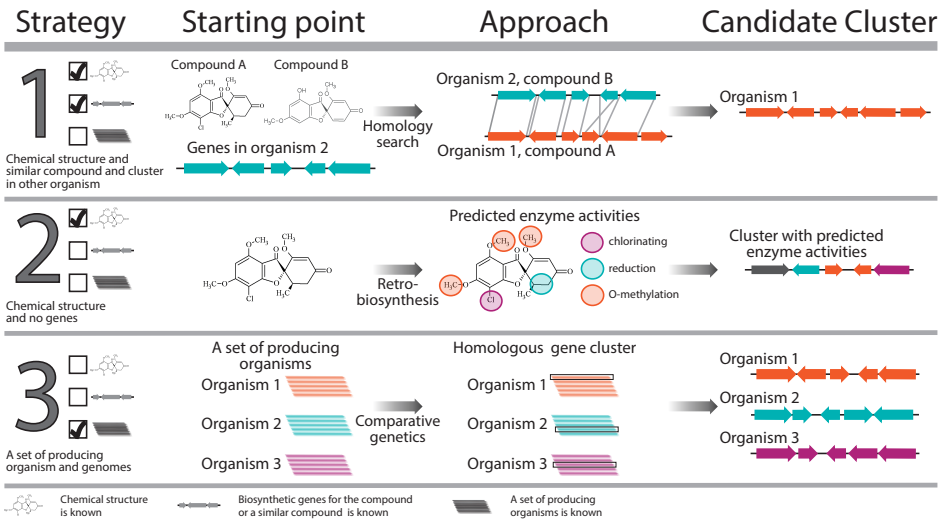


Figure 2: Three strategies for linking secondary metabolite to clusters using whole genome sequences. 1) Homology search – starting with a known compound produced by organisms A and the same or similar compound produced by organism B where the cluster has been identified it is possible to use the known cluster from organism B to search for a similar cluster in the genome of organism A and thereby identify the cluster of interest. 2) Retro-biosynthesis – starting with a known compound but no similar clusters identified it is possible the predict the enzyme activities needed to produce the compound (backbone and tailoring enzymes) and from these predictions find putative clusters matching the requirements in the genome. 3) Comparative genomic – starting with a set of organisms where some produces the compound of interest and some do not it is possible to identify homologous gene clusters in the producing and filter based on no homologs in the non producing and thereby identify candidate gene clusters.

wheat pathogen *Zymoseptoria tritici* using MultiGene Basic Local Alignment Search Tools (BLAST) [69] to identify known clusters. This strategy revealed a cluster similar to the ferrichrome-A biosynthetic locus from a maize pathogen, *Ustilago maydis*. Putative clusters for carotenoid/opsin, an epipolythiodioxopiperazine, fumonisin and AM-toxin were also identified [24]. After whole genome sequencing of *Penicillium griseofulvum* various cluster products were predicted based on homology to known clusters from the MIBiG database [68]. This revealed that *Penicillium griseofulvum* has putative gene clusters for patulin, roquefortine C / meleagrin, griseofulvin, penicillin, cyclopiazonic acid, yanuthone D and chanoclavine I [7]. Similarly, the genome of *Aspergillus ustus* a rare human pathogen was searched for known clusters and several commonly known aspergillus SM clusters were identified; monodictyphenone, sterigmatocystin, emericellamide, ferricrocin and asperthecin. In addition a putative cluster for viridicatumtoxin not previously identified in Aspergillus species was identified in *A. ustus* [83]. A homology search approach (based on ClusterBLAST module embedded within FungiSMASH [15]) was also applied in the investigation of the lichen, *Cladonia uncialis*. This revealed genes homologs to the lichen metabolite grayanic acid from

C. grayi in addition to clusters likely encoding fungal SMs not identified in lichens before such as patulin and betaenones AC [13].

The immediate advantage of this method is that it can be used on all species as long as there is a genome sequence. The disadvantage is that it requires both a genome sequence and a similar cluster already characterized, and it only finds known clusters or derivatives but not truly new SMs. Lastly, the results are putative and requires some experimental verification. Homology search is an extremely powerful and highly employed approach for coupling clusters and SMs especially in the investigation of newly sequenced genomes. As the number of whole genome sequences increases and the algorithms aiding in the predictions are improved, this method is likely to expand.

3.2. Retro Biosynthesis

The second strategy seen on Figure 2 is based on retro-biosynthesis, here the secondary metabolite (SM) is known and the chemical structure of it but nothing else. Based on the chemical structure and knowledge of secondary metabolite biosynthesis it is possible to deduce what enzyme activities are needed to produce the SM which can be used to identify a putative gene cluster matching the requirements.

The anticancer lipopeptide, scopularide A, is produced by a marine derived *Scopulariopsis brevicaulis* and the chemical structure consists of a reduced carbon chain coupled to five amino acids. The SM is structurally related to emericellamide A and W493-B from *Aspergillus nidulans* and *Fusarium pseudograminearum* respectively. After the sequencing of *S. brevicaulis*, Lukassen et al. wanted to identify the cluster responsible for scopularide A production in order to optimize the production. This was done primarily based on retro-biosynthetic approached supported by homologous comparisons. By combining the knowledge of the structure with predicted secondary metabolite gene clusters it was possible to identify genes encoding the SM, a NRPS with five modules and a reducing PKS. The identified genes also showed homology to the clusters for the structurally related SMs emericellamide A and W493-B. The putative cluster included a predicted TF. To further support the prediction and to improve the production of scopularide A the TF was overexpressed which significantly increased the production of scopularide A thus indirectly verifying the prediction [63].

A retro-biosynthesis based approach was also used in the identification of the putative usnic acid cluster in the lichen fungal partner of *Cladonia uncialis* [1]. After de novo sequencing of *C. uncialis*, the genome was mined for PKS genes. From the structure of usnic acid and an earlier labelling experiment, it was suggested that usinic acid biosynthesis requires a non-reducing PKS including a methylation domain and a terminal Claisen cyclase (CLC) domain plus an oxidative tailoring enzyme, most likely a cytochrome P450. Based on this information, the predicted PKS clusters were screened, and only one matched the requirements. Transcriptional analysis of the genes was performed under conditions where only usnic acid was produced. This confirmed that the identified genes were transcriptionally active which further supports the predictions.

Khater et al. have attempted to develop a computational protocol based on the concept of retro-biosynthesis to reconstruct biosynthetic pathways of polyketides and nonribosomal

peptides [56]. The aim is to predict the enzymes and the gene functions involved in the biosynthesis of a certain SM and thus be able to predict and identify the responsible biosynthetic gene cluster in an automated manner. The developed approach was tested based on 78 experimentally characterized secondary metabolites (51 PKS/HYBRID, 27 NRPS), here it was able to predict 37% correctly, 13% with minor errors, 24% partially correct and 26% incorrectly. The predictive methods are still in the early stage and needs more knowledge of the mechanisms behind the secondary metabolite production and development but it has the potential to become a very powerful tool in the future.

3.3. Comparative Genomics

Comparative genomics is the third approach identified in our metaanalysis, Figure 2. This approach requires the whole genome sequences of several species where the differences or similarities can be used to identify a specific gene cluster. The species used could for instance be distantly related species producing the same secondary metabolite (SM) or closely related species sharing high degree of secondary metabolism but not the SM of interest.

The identification of the viridicatumtoxin and griseofulvin gene clusters in *Penicillium aethiopicum* was accomplished using such comparative genomics strategies: *P. aethiopicum* was sequenced and compared to the *Penicillium chrysogenum* genome which is a closely related species but it does not produce the SMs of interest. This way 9 out of 30 predicted PKSs could be ruled out due to homology between them. To further narrow it down retro-biosynthetic methods was used to identify the most likely kind of PKS and to the check that the surrounding genes match the expected tailoring enzymes. This way candidate clusters for both SMs were identified and these were verified by gene deletion and RNA silencing [29].

The biosynthetic clusters of the (+)/(-)-notoamide, paraherquamide and malbrancheamide pathways (all based on bicyclo[2.2.2]diazaoctane indole alkaloid core) were identified based on homology search and comparative genomics. The genomes of *A. versicolor* NRRL35600, *P. fellutanum* ATCC20841, and *M. aurantiaca* were sequenced and the (-)notoamide cluster known from *Aspergillus* sp. MF297-2 [34] was used to search for homologs in the newly sequenced species [62]. Comparison of the identified clusters led to the identification of genes responsible in the formation of the bicyclo[2.2.2]diazaoctane core along with specific enzymes responsible for specific differences in the chemical structures [62].

The advantages of using comparative genomics is that no knowledge about the cluster or biosynthesis is required, only a set of genomes to show a specific pattern for the cluster of interest (e.g. the corresponding species produce a SM of interest). Disadvantages includes that whole genome sequences are required as well as knowledge of producer and non-producers of the SM of interest. Caution needs to be applied since non-producers could have silent clusters and therefore a chemical negative should not be taken as an absolute but rather a indication. As more and more species are sequenced, we expect that the comparative analysis will expand and more sophisticated methods will develop.

Comparative genomics strategies are often used with a combination of retro-biosynthesis and/or homology search, again showing that a combination of strategies is most often needed to establish the links between SM and gene clusters.

4. Conclusion

Many strategies have been developed and employed in the quest to link secondary metabolites to their secondary metabolite gene clusters and vice versa. As the technologies and tools continue to evolve and we get a deeper understanding of the mechanisms behind secondary metabolite production the speed and efficiency of linking SM and clusters will increase.

We see three major areas of advancement in the near future. Firstly, molecular tools are developing making it feasible to work with many different species and thereby making it possible to conduct analysis in the native species. We will therefore see more studies from non-model organisms. Secondly, as the price of de novo synthesis of DNA is rapidly decreasing heterologous expression of silent clusters will become easier, opening for larger high throughput screening studies. Thirdly, with the increasing number of whole genome sequences and knowledge more compartive genomics approaches will be used and advanced bioinformtaic tools will emerge making more accurate and more advanced predictions.

References

[1] Abdel-Hameed, M., Bertrand, R.L., Piercey-Normore, M.D., Sorensen, J.L.. Putative identification of the usnic acid biosynthetic gene cluster by de novo whole-genome sequencing of a lichen-forming fungus. Fungal Biology 2016;120(3):306–316.

[2] Agger, S., Lopez-Gallego, F., Schmidt-Dannert, C.. Diversity of sesquiterpene synthases in the basidiomycete Coprinus cinereus. Molecular Microbiology 2009;72(5):1181–1195.

[3] Alberti, F., Foster, G.D., Bailey, A.M.. Natural products from filamentous fungi and production by heterologous expression. Applied Microbiology and Biotechnology 2017;101(2):493–500.

[4] Alves, P.C., Hartmann, D.O., Núñez, O., Martins, I., Gomes, T.L., Garcia, H., Galceran, M.T., Hampson, R., Becker, J.D., Pereira, C.S.. Transcriptomic and metabolomic profiling of ionic liquid stimuli unveils enhanced secondary metabolism in Aspergillus nidulans. BMC Genomics 2016;17:284.

[5] Anyaogu, D.C., Mortensen, U.H.. Heterologous production of fungal secondary metabolites in Aspergilli. Frontiers in Microbiology 2015;6:77.

[6] Bailey, A.M., Alberti, F., Kilaru, S., Collins, C.M., De Mattos-Shipley, K., Hartley, A.J., Hayes, P., Griffin, A., Lazarus, C.M., Cox, R.J., Willis, C.L., O'Dwyer, K., Spence, D.W., Foster, G.D.. Identification and manipulation of the pleuromutilin gene cluster from Clitopilus passeckerianus for increased rapid antibiotic production. Scientific Reports 2016;6.

[7] Banani, H., Marcet-Houben, M., Ballester, A.R., Abbruscato, P., González-Candelas, L., Gabaldón, T., Spadaro, D.. Genome sequencing and secondary metabolism of the postharvest pathogen Penicillium griseofulvum. BMC Genomics 2016;17(19).

[8] Bayram, O., Krappmann, S., Ni, M., Bok, J.W., Helmstaedt, K., Valerius, O., Braus-Stromeyer, S., Kwon, N.J., Keller, N.P., Yu, J.H., Braus, G.H.. VelB/VeA/LaeA complex coordinates light signal with fungal development and secondary metabolism. Science 2008;320(5882):1504–6.

[9] Bedford, D.J., Schweizer, E., Hopwood, D.A., Khosla, C.. Expression of a Functional Fungal Polyketide Synthase in the Bacterium Streptomyces coelicolor A3(2). JOURNAL OF BACTERIOLOGY 1995;177(15):4544–4548.

[10] Bennett, J.W., Klich, M.. Mycotoxins. Clinical Microbiology Reviews 2003;16(3):497–516.

[11] Bergmann, S., Funk, A.N., Scherlach, K., Schroeckh, V., Shelest, E., Horn, U., Hertweck, C., Brakhage, A.A.. Activation of a silent fungal polyketide biosynthesis pathway through regulatory cross talk with a cryptic nonribosomal peptide synthetase gene cluster. Applied and Environmental Microbiology 2010;76(24):8143–8149.

[12] Bergmann, S., Schümann, J., Scherlach, K., Lange, C., Brakhage, A.A., Hertweck, C.. Genomics-driven discovery of PKS-NRPS hybrid metabolites from Aspergillus nidulans. Nature chemical biology 2007;3(4):213–217.

[13] Bertrand, R.L., Abdel-Hameed, M., Sorensen, J.L.. Lichen Biosynthetic Gene Clusters Part II: Homology Mapping Suggests a Functional Diversity. Journal of Natural Products 2018;81(4):732–748.

[14] Bi, Q., Wu, D., Zhu, X., Turgeon, B.G.. Cochliobolus heterostrophus Llm1 - A Lae1-like methyltransferase regulates T-toxin production, virulence, and development. Fungal Genetics and Biology 2013;51:21–33.

[15] Blin, K., Wolf, T., Chevrette, M.G., Lu, X., Schwalen, C.J., Kautsar, S.A., Suarez Duran, H.G., De Los Santos, E.L., Kim, H.U., Nave, M., Dickschat, J.S., Mitchell, D.A., Shelest, E., Breitling, R., Takano, E., Lee, S.Y., Weber, T., Medema, M.H.. AntiSMASH 4.0 - improvements in chemistry prediction and gene cluster boundary identification. Nucleic Acids Research 2017;45(Web server issue):W36–W41.

[16] Bok, J.W., Chiang, Y.M., Szewczyk, E., Reyes-Dominguez, Y., Davidson, A.D., Sanchez, J.F., Lo, H.C., Watanabe, K., Strauss, J., Oakley, B.R., Wang, C.C.C., Keller, N.P.. Chromatin-level regulation of biosynthetic gene clusters. Nature chemical biology 2009;5(7):462–464.

[17] Bok, J.W., Hoffmeister, D., Maggio-Hall, L.A., Murillo, R., Glasner, J.D., Keller, N.P.. Genomic Mining for Aspergillus Natural Products. Chemistry & Biology 2006;13:31–37.

[18] Bok, J.W., Keller, N.P.. LaeA, a regulator of secondary metabolism in Aspergillus spp. Eukaryotic cell 2004;3(2):527–35.

[19] Bond, C., Tang, Y., Li, L.. Saccharomyces cerevisiae as a tool for mining, studying and engineering fungal polyketide synthases. Fungal Genetics and Biology 2016;89:52–61.

[20] Brakhage, A.A.. Regulation of fungal secondary metabolism. 2013.

[21] Bromann, K., Toivari, M., Viljanen, K., Vuoristo, A., Ruohonen, L., Nakari-Setä lä, T., Henrique Goldman, G.. Identification and Characterization of a Novel Diterpene Gene Cluster in Aspergillus nidulans. PLoS ONE 2012;7(4):e35450.

[22] Butchko, R.A.E., Brown, D.W., Busman, M., Tudzynski, B., Wiemann, P.. Lae1 regulates expression of multiple secondary metabolite gene clusters in Fusarium verticillioides. Fungal Genetics and Biology 2012;49:602–612.

[23] Cacho, R.A., Jiang, W., Chooi, Y.H., Walsh, C.T., Tang, Y.. Identification and characterization of the echinocandin b biosynthetic gene cluster from Emericella rugulosa NRRL 11440. Journal of the American Chemical Society 2012;134(40):16781–16790.

[24] Cairns, T., Meyer, V.. In silico prediction and characterization of secondary metabolite biosynthetic gene clusters in the wheat pathogen Zymoseptoria tritici. BMC Genomics 2017;18(1).

[25] Chang, P.K., Horn, B.W., Dorner, J.W.. Clustered genes involved in cyclopiazonic acid production are next to the aflatoxin biosynthesis gene cluster in Aspergillus flavus. Fungal Genetics and Biology 2009;46(2):176–182.

[26] Chiang, Y.M., Meyer, K.M., Praseuth, M., Baker, S.E., Bruno, K.S., Wang, C.C.C.. Characterization of a polyketide synthase in Aspergillus niger whose product is a precursor for both dihydroxynaphthalene (DHN) melanin and naphtho-γ-pyrone. Fungal Genetics and Biology 2011;48(4):430–437.

[27] Chiang, Y.M., Oakley, E., Ahuja, M., Entwistle, R., Schultz, A., Chang, S.L., Sung, C.T., Wang, C.C.C., Oakley, B.R.. An Efficient System for Heterologous Expression of Secondary Metabolite Genes in Aspergillus nidulans. Journal of the American Chemical Society 2013;135(20):7720–7731.

[28] Chiang, Y.M., Szewczyk, E., Davidson, A.D., Keller, N., Oakley, B.R., Wang, C.C.. A gene cluster containing two fungal polyketide synthases encodes the biosynthetic pathway for a polyketide, asperfuranone, in aspergillus nidulans. Journal of the American Chemical Society 2009;131(8):2965–2970.

[29] Chooi, Y.H., Cacho, R., Tang, Y.. Identification of the Viridicatumtoxin and Griseofulvin Gene Clusters from Penicillium aethiopicum. Chemistry and Biology 2010;17(5):483–494.

[30] Clevenger, K.D., Bok, J.W., Ye, R., Miley, G.P., Verdan, M.H., Velk, T., Chen, C., Yang, K.,

Robey, M.T., Gao, P., Lamprecht, M., Thomas, P.M., Islam, M.N., Palmer, J.M., Wu, C.C., Keller, N.P., Kelleher, N.L.. A scalable platform to identify fungal secondary metabolites and their gene clusters. Nature chemical biology 2017;13(8):895.

[31] Cochrane, R.V., Sanichar, R., Lambkin, G.R., Reiz, B., Xu, W., Tang, Y., Vederas, J.C.. Production of New Cladosporin Analogues by Reconstitution of the Polyketide Synthases Responsible for the Biosynthesis of this Antimalarial Agent. Angewandte Chemie - International Edition 2016;55(2):664–668.

[32] Davison, J., al Fahad, A., Cai, M., Song, Z., Yehia, S.Y., Lazarus, C.M., Bailey, A.M., Simpson, T.J., Cox, R.J.. Genetic, molecular, and biochemical basis of fungal tropolone biosynthesis. Proceedings of the National Academy of Sciences 2012;109(20):7642–7647.

[33] De Souza, C.P., Hashmi, S.B., Osmani, A.H., Andrews, P., Ringelberg, C.S., Dunlap, J.C., Osmani, S.A., Yu, J.H.. Functional Analysis of the Aspergillus nidulans Kinome. PLoS ONE 2013;8(3):e58008.

[34] Ding, Y., Wet, J.R.D., Cavalcoli, J., Li, S., Greshock, T.J., Miller, K.A., Finefield, J.M., Sunderhaus, J.D., McAfoos, T.J., Tsukamoto, S., Williams, R.M., Sherman, D.H.. Genome-based characterization of two prenylation steps in the assembly of the stephacidin and notoamide anticancer agents in a marine-derived aspergillus sp. Journal of the American Chemical Society 2010;132(36):12733–12740.

[35] Dowzer, C.E., Kelly, J.M.. Analysis of the creA gene, a regulator of carbon catabolite repression in Aspergillus nidulans. Molecular and cellular biology 1991;11(11):5701–9.

[36] Esperón, P., Scazzocchio, C., Paulino, M.. In vitro and in silico analysis of the Aspergillus nidulans DNACreA repressor interactions. Journal of Biomolecular Structure and Dynamics 2014;32(12):2033–2041.

[37] Espeso, E.A., Peñalva, M.A.. Three binding sites for the Aspergillus nidulans PacC zinc-finger transcription factor are necessary and sufficient for regulation by ambient pH of the isopenicillin N synthase gene promoter. Journal of Biological Chemistry 1996;271(46):28825–28830.

[38] Farzadfard, F., Perli, S.D., Lu, T.K.. Tunable and Multifunctional Eukaryotic Transcription Factors Based on CRISPR/Cas. ACS Synthetic Biology 2013;2:604–613.

[39] Fisch, K.M., Gillaspy, A.F., Gipson, M., Henrikson, J.C., Hoover, A.R., Jackson, L., Najar, F.Z., Wägele, H., Cichewicz, R.H.. Chemical induction of silent biosynthetic pathway transcription in aspergillus niger. Journal of Industrial Microbiology and Biotechnology 2009;36(9):1199–1213.

[40] Fujii, R., Minami, A., Tsukagoshi, T., Sato, N., Sahara, T., Ohgiya, S., Gomi, K., Oikawa, H.. Total Biosynthesis of Diterpene Aphidicolin, a Specific Inhibitor of DNA Polymerase α: Heterologous Expression of Four Biosynthetic Genes in Aspergillus oryzae. Bioscience, Biotechnology, and Biochemistry 2011;75(9):1813–1817.

[41] Gaffoor, I., Brown, D.W., Plattner, R., Proctor, R.H., Qi, W., Trail, F.. Functional analysis of the polyketide synthase genes in the filamentous fungus Gibberella zeae (anamorph Fusarium graminearum). Eukaryot Cell 2005;4(11):1926–1933.

[42] Gallo, A., Knox, B.P., Bruno, K.S., Solfrizzo, M., Baker, S.E., Perrone, G.. Identification and characterization of the polyketide synthase involved in ochratoxin A biosynthesis in Aspergillus carbonarius. International Journal of Food Microbiology 2014;179:10–17.

[43] Gao, X., Chooi, Y.H., Ames, B.D., Wang, P., Walsh, C.T., Tang, Y.. Fungal indole alkaloid biosynthesis: Genetic and biochemical investigation of the tryptoquialanine pathway in penicillium aethiopicum. Journal of the American Chemical Society 2011;133(8):2729–2741.

[44] Gao, X., Wang, P., Tang, Y.. Engineered polyketide biosynthesis and biocatalysis in Escherichia coli. Applied Microbiology and Biotechnology 2010;88(6):1233–1242.

[45] Gressler, M., Zaehle, C., Scherlach, K., Hertweck, C., Brock, M.. Multifactorial induction of an orphan PKS-NRPS gene cluster in Aspergillus terreus. Chemistry and Biology 2011;18(2):198–209.

[46] Harvey, C.J., Tang, M., Schlecht, U., Horecka, J., Fischer, C.R., Lin, H.C., Li, J., Naughton, B., Cherry, J., Miranda, M., Li, Y.F., Chu, A.M., Hennessy, J.R., Vandova, G.A., Inglis, D., Aiyar, R.S., Steinmetz, L.M., Davis, R.W., Medema, M.H., Sattely, E., Khosla, C., Onge, R.P., Tang,

Y., Hillenmeyer, M.E.. HEx: A heterologous expression platform for the discovery of fungal natural products. Science Advances 2018;4(4):eaar5459.

[47] Hautbergue, T., Jamin, E.L., Debrauwer, L., Puel, O., Oswald, I.P.. From genomics to metabolomics, moving toward an integrated strategy for the discovery of fungal secondary metabolites. Natural product reports 2018;35(2):147–173.

[48] Heneghan, M.N., Yakasai, A.A., Halo, L.M., Song, Z., Bailey, A.M., Simpson, T.J., Cox, R.J., Lazarus, C.M.. First Heterologous Reconstruction of a Complete Functional Fungal Biosynthetic Multigene Cluster. ChemBioChem 2010;11:1508–1512.

[49] Ishiuchi, K., Nakazawa, T., Ookuma, T., Sugimoto, S., Sato, M., Tsunematsu, Y., Ishikawa, N., Noguchi, H., Hotta, K., Moriya, H., Watanabe, K.. Establishing a New Methodology for Genome Mining and Biosynthesis of Polyketides and Peptides through Yeast Molecular Genetics. ChemBioChem 2012;13(6):846–854.

[50] Itoh, T., Tokunaga, K., Matsuda, Y., Fujii, I., Abe, I., Ebizuka, Y., Kushiro, T.. Reconstitution of a fungal meroterpenoid biosynthesis reveals the involvement of a novel family of terpene cyclases. Nature Chemistry 2010;2(10):858–864.

[51] Jain, S., Keller, N.. Insights to fungal biology through LaeA sleuthing. Fungal Biology Reviews 2013;27:51–59.

[52] Janevska, S., Tudzynski, B.. Secondary metabolism in Fusarium fujikuroi: strategies to unravel the function of biosynthetic pathways. Applied Microbiology and Biotechnology 2018;102:615–630.

[53] Kealey, J.T., Liu, L., Santi, D.V., Betlach, M.C., Barr, P.J.. Production of a polyketide natural product in nonpolyketide- producing prokaryotic and eukaryotic hosts. Proceedings of the National Academy of Sciences 1998;95:505–509.

[54] Keller, N.P., Turner, G., Bennett, J.W.. Fungal secondary metabolism from biochemistry to genomics. Nature Reviews Microbiology 2005;3(12):937–947.

[55] Khaldi, N., Seifuddin, F.T., Turner, G., Haft, D., Nierman, W.C., Wolfe, K.H., Fedorova, N.D.. SMURF: Genomic mapping of fungal secondary metabolite clusters. Fungal genetics and biology : FG & B 2010;47(9):736–41.

[56] Khater, S., Anand, S., Mohanty, D.. In silico methods for linking genes and secondary metabolites: The way forward. Synthetic and Systems Biotechnology 2016;1:80–88.

[57] Kjærbølling, I., Vesth, T.C., Frisvad, J.C., Nybo, J.L., Theobald, S., Kuo, A., Bowyer, P., Matsuda, Y., Mondo, S., Lyhne, E.K., Kogle, M.E., Clum, A., Lipzen, A., Salamov, A., Ngan, C.Y., Daum, C., Chiniquy, J., Barry, K., LaButti, K., Haridas, S., Simmons, B.A., Magnuson, J.K., Mortensen, U.H., Larsen, T.O., Grigoriev, I.V., Baker, S.E., Andersen, M.R.. Linking secondary metabolites to gene clusters through genome sequencing of six diverse Aspergillus species. Proceedings of the National Academy of Sciences 2018;115(4):E753–E761.

[58] Kupfer, D.M., Drabenstot, S.D., Buchanan, K.L., Lai, H., Zhu, H., Dyer, D.W., Roe, B.A., Murphy, J.W.. Introns and splicing elements of five diverse fungi. Eukaryotic Cell 2004;3(5):1088–1100.

[59] Kwon-Chung, K.J., Sugui, J.A.. What do we know about the role of gliotoxin in the pathobiology of Aspergillus fumigatus? Medical Mycology 2009;47(SUPPL. 1):S97–103.

[60] Lambalot, R.H., Gehring, A.M., Flugel, R.S., Zuber, P., LaCelle, M., Marahiel, M.A., Reid, R., Khosla, C., Walsh, C.T.. A new enzyme superfamily - The phosphopantetheinyl transferases. Chemistry and Biology 1996;3(11):923–936.

[61] Lazarus, C.M., Williams, K., Bailey, A.M.. Reconstructing fungal natural product biosynthetic pathways. Natural Product Reports 2014;31:1339.

[62] Li, S., Srinivasan, K., Tran, H., Yu, F., Finefield, J.M., Sunderhaus, J.D., McAfoos, T.J., Tsukamoto, S., Williams, R.M., Sherman, D.H.. Comparative analysis of the biosynthetic systems for fungal bicyclo[2.2.2]diazaoctane indole alkaloids: The (+)/(-)-notoamide, paraherquamide and malbrancheamide pathways. MedChemComm 2012;3(8):987–996.

[63] Lukassen, M., Saei, W., Sondergaard, T., Tamminen, A., Kumar, A., Kempken, F., Wiebe, M., Sørensen, J.. Identification of the Scopularide Biosynthetic Gene Cluster in Scopulariopsis brevicaulis.

Marine Drugs 2015;13(7):4331–4343.

[64] Ma, S.M., Li, J.W., Choi, J.W., Zhou, H., Lee, K.K., Moorthie, V.A., Xie, X., Kealey, J.T., Da Silva, N.A., Vederas, J.C., Tang, Y.. Complete reconstitution of a highly reducing iterative polyketide synthase. Science 2009;326(5952):589–592.

[65] Ma, S.M., Zhan, J., Watanabe, K., Xie, X., Zhang, W., Wang, C.C., Tang, Y.. Enzymatic synthesis of aromatic polyketides using PKS4 from Gibberella fujikuroi. Journal of the American Chemical Society 2007;129(35):10642–10643.

[66] Malz, S., Grell, M.N., Thrane, C., Maier, F.J., Rosager, P., Felk, A., Albertsen, K.S., Salomon, S., Bohn, L., Schäfer, W., Giese, H.. Identification of a gene cluster responsible for the biosynthesis of aurofusarin in the Fusarium graminearum species complex. Fungal Genetics and Biology 2005;42(5):420–433.

[67] Matsuda, Y., Bai, T., Phippen, C.B., Nødvig, C.S., Kjærbølling, I., Vesth, T.C., Andersen, M.R., Mortensen, U.H., Gotfredsen, C.H., Abe, I., Larsen, T.O.. Novofumigatonin biosynthesis involves a non-heme iron-dependent endoperoxide isomerase for orthoester formation. Nature Communications 2018;9(1):2587.

[68] Medema, M.H., Kottmann, R., Yilmaz, P., Cummings, M., Biggins, J.B., Blin, K., de Bruijn, I., Chooi, Y.H., Claesen, J., Coates, R.C., Cruz-Morales, P., Duddela, S., Düsterhus, S., Edwards, D.J., Fewer, D.P., Garg, N., Geiger, C., Gomez-Escribano, J.P., Greule, A., Hadjithomas, M., Haines, A.S., Helfrich, E.J.N., Hillwig, M.L., Ishida, K., Jones, A.C., Jones, C.S., Jungmann, K., Kegler, C., Kim, H.U., Kötter, P., Krug, D., Masschelein, J., Melnik, A.V., Mantovani, S.M., Monroe, E.A., Moore, M., Moss, N., Nützmann, H.W., Pan, G., Pati, A., Petras, D., Reen, F.J., Rosconi, F., Rui, Z., Tian, Z., Tobias, N.J., Tsunematsu, Y., Wiemann, P., Wyckoff, E., Yan, X., Yim, G., Yu, F., Xie, Y., Aigle, B., Apel, A.K., Balibar, C.J., Balskus, E.P., Barona-Gómez, F., Bechthold, A., Bode, H.B., Borriss, R., Brady, S.F., Brakhage, A.A., Caffrey, P., Cheng, Y.Q., Clardy, J., Cox, R.J., De Mot, R., Donadio, S., Donia, M.S., van der Donk, W.A., Dorrestein, P.C., Doyle, S., Driessen, A.J.M., Ehling-Schulz, M., Entian, K.D., Fischbach, M.A., Gerwick, L., Gerwick, W.H., Gross, H., Gust, B., Hertweck, C., Höfte, M., Jensen, S.E., Ju, J., Katz, L., Kaysser, L., Klassen, J.L., Keller, N.P., Kormanec, J., Kuipers, O.P., Kuzuyama, T., Kyrpides, N.C., Kwon, H.J., Lautru, S., Lavigne, R., Lee, C.Y., Linquan, B., Liu, X., Liu, W., Luzhetskyy, A., Mahmud, T., Mast, Y., Méndez, C., Metsä-Ketelä, M., Micklefield, J., Mitchell, D.A., Moore, B.S., Moreira, L.M., Müller, R., Neilan, B.A., Nett, M., Nielsen, J., O'Gara, F., Oikawa, H., Osbourn, A., Osburne, M.S., Ostash, B., Payne, S.M., Pernodet, J.L., Petricek, M., Piel, J., Ploux, O., Raaijmakers, J.M., Salas, J.A., Schmitt, E.K., Scott, B., Seipke, R.F., Shen, B., Sherman, D.H., Sivonen, K., Smanski, M.J., Sosio, M., Stegmann, E., Süssmuth, R.D., Tahlan, K., Thomas, C.M., Tang, Y., Truman, A.W., Viaud, M., Walton, J.D., Walsh, C.T., Weber, T., van Wezel, G.P., Wilkinson, B., Willey, J.M., Wohlleben, W., Wright, G.D., Ziemert, N., Zhang, C., Zotchev, S.B., Breitling, R., Takano, E., Glöckner, F.O.. The Minimum Information about a Biosynthetic Gene cluster (MIBiG) specification. Nature chemical biology 2015;11(9):625–631.

[69] Medema, M.H., Takano, E., Breitling, R.. Detecting sequence homology at the gene cluster level with multigeneblast. Molecular Biology and Evolution 2013;30(5):1218–1223.

[70] Mihlan, M., Homann, V., Liu, T.W.D., Tudzynski, B.. AREA directly mediates nitrogen regulation of gibberellin biosynthesis in Gibberella fujikuroi, but its activity is not affected by NMR. Molecular Microbiology 2003;47(4):975–991.

[71] Mootz, H.D., Schörgendorfer, K., Marahiel, M.A.. Functional characterization of 4-phosphopantetheinyl transferase genes of bacterial and fungal origin by complementation of Saccharomyces cerevisiae lys5. FEMS Microbiology Letters 2002;213(1):51–57.

[72] Munawar, A., Marshall, J.W., Cox, R.J., Bailey, A.M., Lazarus, C.M.. Isolation and Characterisation of a Ferrirhodin Synthetase Gene from the Sugarcane Pathogen Fusarium sacchari. ChemBioChem 2013;14(3):388–394.

[73] Mutka, S.C., Bondi, S.M., Carney, J.R., Da Silva, N.A., Kealey, J.T.. Metabolic pathway engineering for complex polyketide biosynthesis in Saccharomyces cerevisiae. FEMS Yeast Research

2006;6(1):40–47.

[74] Neubauer, L., Dopstadt, J., Humpf, H.U., Tudzynski, P.. Identification and characterization of the ergochrome gene cluster in the plant pathogenic fungus Claviceps purpurea. Fungal Biology and Biotechnology 2016;3(1):2.

[75] Nielsen, M.L., Nielsen, J.B., Rank, C., Klejnstrup, M.L., Holm, D.K., Brogaard, K.H., Hansen, B.G., Frisvad, J.C., Larsen, T.O., Mortensen, U.H.. A genome-wide polyketide synthase deletion library uncovers novel genetic links to polyketides and meroterpenoids in Aspergillus nidulans. FEMS microbiology letters 2011;321(2):157–166.

[76] Nielsen, M.T., Nielsen, J.B., Anyaogu, D.C., Holm, D.K., Nielsen, K.F., Larsen, T.O., Mortensen, U.H.. Heterologous Reconstitution of the Intact Geodin Gene Cluster in Aspergillus nidulans through a Simple and Versatile PCR Based Approach. PLoS ONE 2013;8(8):72871.

[77] Nierman, W.C., Pain, A., Anderson, M.J., Wortman, J.R., Kim, H.S., Arroyo, J., Berriman, M., Abe, K., Archer, D.B., Bermejo, C., Bennett, J., Bowyer, P., Chen, D., Collins, M., Coulsen, R., Davies, R., Dyer, P.S., Farman, M., Fedorova, N., Fedorova, N., Feldblyum, T.V., Fischer, R., Fosker, N., Fraser, A., García, J.L., García, M.J., Goble, A., Goldman, G.H., Gomi, K., Griffith-Jones, S., Gwilliam, R., Haas, B., Haas, H., Harris, D., Horiuchi, H., Huang, J., Humphray, S., Jiménez, J., Keller, N., Khouri, H., Kitamoto, K., Kobayashi, T., Konzack, S., Kulkarni, R., Kumagai, T., Lafton, A., Latgé, J.P., Li, W., Lord, A., Lu, C., Majoros, W.H., May, G.S., Miller, B.L., Mohamoud, Y., Molina, M., Monod, M., Mouyna, I., Mulligan, S., Murphy, L., O'Neil, S., Paulsen, I., Peñalva, M.A., Pertea, M., Price, C., Pritchard, B.L., Quail, M.A., Rabbinowitsch, E., Rawlins, N., Rajandream, M.A., Reichard, U., Renauld, H., Robson, G.D., Rodriguez De Cordoba, S., Rodríguez-Peña, J.M., Ronning, C.M., Rutter, S., Salzberg, S.L., Sanchez, M., Sánchez-Ferrero, J.C., Saunders, D., Seeger, K., Squares, R., Squares, S., Takeuchi, M., Tekaia, F., Turner, G., Vazquez De Aldana, C.R., Weidman, J., White, O., Woodward, J., Yu, J.H., Fraser, C., Galagan, J.E., Asai, K., Machida, M., Hall, N., Barrell, B., Denning, D.W.. Genomic sequence of the pathogenic and allergenic filamentous fungus Aspergillus fumigatus. Nature 2005;438(7071):1151–1156.

[78] Nødvig, C.S., Blaesbjerg, J., Engelhard, M., Hasbro, U., Nødvig, C.S., Nielsen, J.B., Kogle, M.E., Mortensen, U.H.. A CRISPR-Cas9 System for Genetic Engineering of Filamentous Fungi. Plos One 2015;10(7):e0133085.

[79] Osbourn, A.. Secondary metabolic gene clusters: Evolutionary toolkits for chemical innovation. Trends in Genetics 2010;26(10):449–457.

[80] Payne, G.A., Nierman, W.C., Wortman, J.R., Pritchard, B.L., Brown, D., Dean, R.A., Bhatnagar, D., Cleveland, T.E., Machida, M., Yu, J.. Whole genome comparison of Aspergillus flavus and A. oryzae. Medical Mycology 2006;44:S9–S11.

[81] Perrin, R.M., Fedorova, N.D., Bok, J.W., Cramer, R.A., Wortman, J.R., Kim, H.S., Nierman, W.C., Keller, N.P.. Transcriptional Regulation of Chemical Diversity in Aspergillus fumigatus by LaeA. PLoS Pathogens 2007;3(4):e50.

[82] Pfeifer, B.A., Khosla, C.. Biosynthesis of Polyketides in Heterologous Hosts. Applied and Environmental Microbiology 2001;65(1):106–118.

[83] Pi, B., Yu, D., Dai, F., Song, X., Zhu, C., Li, H., Yu, Y.. A genomics based discovery of secondary metabolite biosynthetic gene clusters in Aspergillus ustus. PLoS ONE 2015;10(2).

[84] Proctor, R.H., Desjardins, A.E., Plattner, R.D., Hohn, T.M.. A polyketide synthase gene required for biosynthesis of fumonisin mycotoxins in Gibberella fujikuroi mating population A. Fungal Genetics and Biology 1999;27(1):100–112.

[85] Reeves, C.D., Hu, Z., Reid, R., Kealey, J.T.. Genes for the Biosynthesis of the Fungal Polyketides Hypothemycin from Hypomyces subiculosus and Radicicol from Pochonia chlamydosporia. APPLIED AND ENVIRONMENTAL MICROBIOLOGY 2008;74(16):5121–5129.

[86] Regueira, T.B., Kildegaard, K.R., Hansen, B.G., Mortensen, U.H., Hertweck, C., Nielsen, J.. Molecular basis for mycophenolic acid biosynthesis in Penicillium brevicompactum. Applied and Environmental Microbiology 2011;77(9):3035–3043.

[87] Robey, M.T., Ye, R., Bok, J.W., Clevenger, K.D., Islam, M.N., Chen, C., Gupta, R., Swyers, M., Wu, E., Gao, P., Thomas, P.M., Wu, C.C., Keller, N.P., Kelleher, N.L.. Identification of the First Diketomorpholine Biosynthetic Pathway Using FAC-MS Technology. ACS Chemical Biology 2018;13(5):1142–1147.

[88] Roze, L.V., Chanda, A., Linz, J.E.. Compartmentalization and molecular traffic in secondary metabolism: A new understanding of established cellular processes. Fungal Genetics and Biology 2011;48(1):35–48.

[89] Rugbjerg, P., Naesby, M., Mortensen, U.H., Frandsen, R.J.N.. Reconstruction of the biosynthetic pathway for the core fungal polyketide scaffold rubrofusarin in Saccharomyces cerevisiae. Microbial Cell Factories 2013;12(1).

[90] Sakai, K., Kinoshita, H., Shimizu, T., Nihira, T.. Construction of a Citrinin Gene Cluster Expression System in Heterologous Aspergillus oryzae. Journal of Bioscience and Bioengineering 2008;106(5):466–472.

[91] Schneider, P., Misiek, M., Hoffmeister, D.. In Vivo and In Vitro Production Options for Fungal Secondary Metabolites. Molecular Phylogenetics 2008;5(2):234–242.

[92] Schumann, J., Hertweck, C.. Cytochalasan biosyn in fungi: Gene cluster analysis & evidence for involvement of a PKS-NRPS hybrid synthase by RNA silencing. J Am Chem Soc 2007;129(31):9564–9565.

[93] Seo, J.A., Proctor, R.H., Plattner, R.D.. Characterization of four clustered and coregulated genes associated with fumonisin biosynthesis in Fusarium verticillioides. Fungal Genetics and Biology 2001;34(3):155–165.

[94] Shwab, E.K., Bok, J.W., Tribus, M., Galehr, J., Graessle, S., Keller, N.P.. Histone deacetylase activity regulates chemical diversity in Aspergillus. Eukaryotic cell 2007;6(9):1656–64.

[95] Siddiqui, M.S., Thodey, K., Trenchard, I., Smolke, C.D.. Advancing secondary metabolite biosynthesis in yeast with synthetic biology tools. FEMS Yeast Research 2012;12(2):144–170.

[96] Smith, D.J., Burnham, M.K., Edwards, J., Earl, A.J., Turner, G.. Smith_1990.pdf. Biotechnology 1990;8:39–41.

[97] Sung, C.T., Chang, S.L., Entwistle, R., Ahn, G., Lin, T.S., Petrova, V., Yeh, H.H., Praseuth, M.B., Chiang, Y.M., Oakley, B.R., Wang, C.C.. Overexpression of a three-gene conidial pigment biosynthetic pathway in Aspergillus nidulans reveals the first NRPS known to acetylate tryptophan. Fungal Genetics and Biology 2017;101:1–6.

[98] Von Bargen, K.W., Niehaus, E.M., Krug, I., Bergander, K., Wü, E.U., Tudzynski, B., Humpf, H.U.. Isolation and Structure Elucidation of Fujikurins AD: Products of the PKS19 Gene Cluster in Fusarium f ujikuroi. Journal of n 2015;78:1809–1815.

[99] Wakefield, J., Hassan, H.M., Jaspars, M., Ebel, R., Rateb, M.E.. Dual Induction of New Microbial Secondary Metabolites by Fungal Bacterial Co-cultivation. Frontiers in Microbiology 2017;8:1284.

[100] Wang, M., Beissner, M., Zhao, H.. Aryl-aldehyde formation in fungal polyketides: Discovery and characterization of a distinct biosynthetic mechanism. Chemistry and Biology 2014;21(2):257–263.

[101] Watanabe, A., Fujii, I., Tsai, H.F., Chang, Y.C., Kwon-Chung, K.J., Ebizuka, Y.. Aspergillus fumigatus alb1 encodes naphthopyrone synthase when expressed in Aspergillus oryzae. FEMS Microbiology letters 2000;192:39–44.

[102] Wawrzyn, G.T., Quin, M.B., Choudhary, S., López-Gallego, F., Schmidt-Dannert, C.. Draft genome of omphalotus olearius provides a predictive framework for sesquiterpenoid natural product biosynthesis in basidiomycota. Chemistry and Biology 2012;19(6):772–783.

[103] Wiemann, P., Sieber, C..M.K., Von Bargen, K.W., Studt, L., Niehaus, E.M., Espino, J.J., Huß, K., Michielse, C.B., Albermann, S., Wagner, D., Bergner, S.V., Connolly, L.R., Fischer, A., Reuter, G., Kleigrewe, K., Bald, T., Wingfield, B.D., Ophir, R., Freeman, S., Hippler, M., Smith, K.M., Brown, D.W., Proctor, R.H., Mü Nsterkö Tter, M., Freitag, M., Humpf, H.U., Gü Ldener, U., Tudzynski, B.. Deciphering the Cryptic Genome: Genome-wide Analyses of the Rice Pathogen Fusarium fujikuroi Reveal Complex Regulation of Secondary Metabolism and Novel Metabolites. PLoS Pathogens 2013;9(6):e1003475.

[104] Williams, R.B., Henrikson, J.C., Hoover, A.R., Lee, A.E., Cichewicz, R.H.. Epigenetic remodeling of the fungal secondary metabolome. Organic & Biomolecular Chemistry 2008;6(11):1895–1897.

[105] Xie, X., Watanabe, K., Wojcicki, W.A., Wang, C.C.C., Tang, Y.. Biosynthesis of Lovastatin Analogs with a Broadly Specific Acyltransferase. Chemistry and Biology 2006;13(11):1161–1169.

[106] Xie, Y., Zhang, W., Li, Y., Wang, M., Cerny, R.L., Shen, Y., Du, L.. Transformation of Fusarium verticillioides with a polyketide gene cluster isolated from a fungal endophyte activates the biosynthesis of fusaric acid. Mycology 2011;2(1):24–29.

[107] Xu, Y., Espinosa-Artiles, P., Schubert, V., ming Xu, Y., Zhang, W., Lin, M., Leslie Gunatilaka, A.A., Süssmuth, R., Molnár, I.. Characterization of the biosynthetic genes for 10,11- dehydrocurvularin, a heat shock response-modulating anticancer fungal polyketide from Aspergillus terreus. Applied and Environmental Microbiology 2013;79(6):2038–2047.

[108] Xu, Y., Orozco, R., Wijeratne, E.M., Gunatilaka, A.A., Stock, S.P., Molnár, I.. Biosynthesis of the Cyclooligomer Depsipeptide Beauvericin, a Virulence Factor of the Entomopathogenic Fungus Beauveria bassiana. Chemistry and Biology 2008;15(9):898–907.

[109] Yaegashi, J., Praseuth, M.B., Tyan, S.W., Sanchez, J.F., Entwistle, R., Chiang, Y.M., Oakley, B.R., Wang, C.C.C.. Molecular Genetic Characterization of the Biosynthesis Cluster of a Prenylated Isoindolinone Alkaloid Aspernidine A in Aspergillus nidulans. ORGANIC LETTERS 2013;15(11):2862–2865.

[110] Yeh, H.H., Ahuja, M., Chiang, Y.M., Oakley, C.E., Moore, S., Yoon, O., Hajovsky, H., Bok, J.W., Keller, N.P., Wang, C.C., Oakley, B.R.. Resistance Gene-Guided Genome Mining: Serial Promoter Exchanges in Aspergillus nidulans Reveal the Biosynthetic Pathway for Fellutamide B, a Proteasome Inhibitor. ACS Chemical Biology 2016;11(8):2275–2284.

[111] Yin, W.B., Chooi, Y.H., Smith, A.R., Cacho, R.A., Hu, Y., White, T.C., Tang, Y.. Discovery of Cryptic Polyketide Metabolites from Dermatophytes Using Heterologous Expression in Aspergillus nidulans. ACS Chemical Biology 2013;2(11):629–634.

[112] Zabala, A.O., Xu, W., Chooi, Y.H., Tang, Y.. Characterization of a silent azaphilone gene cluster from Aspergillus niger ATCC 1015 reveals a hydroxylation-mediated pyran-ring formation. Chemistry & biology 2012;19(8):1049–59.

[113] Zhang, P., Wang, X., Fan, A., Zheng, Y., Liu, X., Wang, S., Zou, H., Oakley, B.R., Keller, N.P., Yin, W.B.. A cryptic pigment biosynthetic pathway uncovered by heterologous expression is essential for conidial development in Pestalotiopsis fici. Molecular Microbiology 2017;105(3):469–483.

3 Exploring genomic diversity and linking compounds to clusters

This chapter includes a paper and a manuscript both investigating *de novo* sequenced species using comparative genomics. The whole genome sequences of diverse and closely related species are used to get new insights and create knowledge-based predictions. An example of this is seen in Paper I, where whole genome sequences are used in a combination with knowledge of secondary metabolite biosynthesis to predict biosynthetic gene clusters of specific compounds. Another example is in Manuscript I where comparative genomics of the important section *Flavi* species are used to gain an understanding of the evolutionary relationship among the species and their capabilities as secondary metabolite and carbohydrate producers. Both Paper I and Manuscript II illustrate the power and versatility of whole genome sequencing and comparative genomics.

3.1 Paper I – Linking compounds to gene clusters through genome sequencing

Paper I is the publication of four high quality *de novo* PacBio sequenced *Aspergillus* species from diverse sections of the *Aspergillus* genus. These genomes will function as reference genomes in the ongoing and future *Aspergillus* comparative work. In addition to these four, two more genomes were sequenced using Illumina technology to support secondary metabolite predictions. Linking secondary metabolites to biosynthetic gene clusters is not a trivial task, however with whole genome sequences at hand, it is possible to use new methods to solve this challenge, as mentioned in section 2.4. Here we show several ways of identifying the biosynthetic gene cluster responsible for producing specific identified compounds one example is the novofumigatonin cluster which was identified here and later the predicted cluster was verified experimentally [1]. This study also compared the opportunistic pathogen *A. fumigatus* with a closely related species *A. novofumigatus* showing that *A. novofumigatus* has the majority of allergens and virulence factors known from *A. fumigatus* suggesting that *A. novofumigatus* has the genetic potential to be pathogenic as well. The analysis showed that *A. novofumigatus* contains many

unique secondary metabolite gene clusters suggesting a big arsenal of potential novel bioactive compounds.

The paper is an excellent example demonstrating various insights and hypothesis that can be gained from comparative genomics studies ranging from biosynthetic cluster predictions to assessing pathogenicity potential.

Paper I was published in Proceedings of the National Academy of Sciences (PNAS) of the United States of America January 9, 2018. The supplementary material can be found in Appendix A.

Linking secondary metabolites to gene clusters through genome sequencing of six diverse *Aspergillus* species

Inge Kjærbølling[a], Tammi C. Vesth[a], Jens C. Frisvad[a], Jane L. Nybo[a], Sebastian Theobald[a], Alan Kuo[b], Paul Bowyer[c], Yudai Matsuda[a], Stephen Mondo[b], Ellen K. Lyhne[a], Martin E. Kogle[a], Alicia Clum[b], Anna Lipzen[b], Asaf Salamov[b], Chew Yee Ngan[b], Chris Daum[b], Jennifer Chiniquy[b], Kerrie Barry[b], Kurt LaButti[b], Sajeet Haridas[b], Blake A. Simmons[d,e], Jon K. Magnuson[d,f], Uffe H. Mortensen[a], Thomas O. Larsen[a], Igor V. Grigoriev[b,g], Scott E. Baker[d,h], and Mikael R. Andersen[a,1]

[a]Department of Biotechnology and Biomedicine, Technical University of Denmark, 2800 Lyngby, Denmark; [b]US Department of Energy Joint Genome Institute, Walnut Creek, CA 94598; [c]Manchester Fungal Infection Group, Institute of Inflammation and Repair, Faculty of Medicine and Human Sciences, University of Manchester, Manchester M13 9PL, United Kingdom; [d]US Department of Energy Joint BioEnergy Institute, Emeryville, CA 94608; [e]Biological Systems and Engineering, Lawrence Berkeley National Laboratory, Berkeley, CA 94720; [f]Energy and Environment Directorate, Pacific Northwest National Laboratory, Richland, WA 99352; [g]Plant and Microbial Biology Department, University of California Berkeley, Berkeley, CA 94720; and [h]Earth and Biological Sciences Directorate, Pacific Northwest National Laboratory, Richland, WA 99352

Edited by Jerrold Meinwald, Cornell University, Ithaca, NY, and approved December 8, 2017 (received for review September 11, 2017)

The fungal genus of *Aspergillus* is highly interesting, containing everything from industrial cell factories, model organisms, and human pathogens. In particular, this group has a prolific production of bioactive secondary metabolites (SMs). In this work, four diverse *Aspergillus* species (*A. campestris*, *A. novofumigatus*, *A. ochraceoroseus*, and *A. steynii*) have been whole-genome PacBio sequenced to provide genetic references in three *Aspergillus* sections. *A. taichungensis* and *A. candidus* also were sequenced for SM elucidation. Thirteen *Aspergillus* genomes were analyzed with comparative genomics to determine phylogeny and genetic diversity, showing that each presented genome contains 15–27% genes not found in other sequenced Aspergilli. In particular, *A. novofumigatus* was compared with the pathogenic species *A. fumigatus*. This suggests that *A. novofumigatus* can produce most of the same allergens, virulence, and pathogenicity factors as *A. fumigatus*, suggesting that *A. novofumigatus* could be as pathogenic as *A. fumigatus*. Furthermore, SMs were linked to gene clusters based on biological and chemical knowledge and analysis, genome sequences, and predictive algorithms. We thus identify putative SM clusters for aflatoxin, chlorflavonin, and ochrindol in *A. ochraceoroseus*, *A. campestris*, and *A. steynii*, respectively, and novofumigatonin, *ent*-cycloechinulin, and *epi*-aszonalenins in *A. novofumigatus*. Our study delivers six fungal genomes, showing the large diversity found in the *Aspergillus* genus; highlights the potential for discovery of beneficial or harmful SMs; and supports reports of *A. novofumigatus* pathogenicity. It also shows how biological, biochemical, and genomic information can be combined to identify genes involved in the biosynthesis of specific SMs.

Aspergillus | fumigatus | comparative genomics | secondary metabolism

The *Aspergillus* genus is a diverse group of fungal species found worldwide in varying habitats. Several species are used in biotechnological industries for the production of enzymes and metabolites (commodity chemicals and pharmaceuticals), and as fermentation agents in food (1). Certain species, such as *A. clavatus* and *A. fumigatus*, are known food spoilers, mycotoxin producers, and opportunistic pathogens (1, 2). To study this diversity, it is important to have reference genomes of high assembly quality in all major clades of the genus. For this purpose, we selected four diverse *Aspergillus* species, *A. campestris*, *A. novofumigatus*, *A. ochraceoroseus*, and *A. steynii*, representing four phylogenetically very different sections in *Aspergillus*, for high-quality PacBio sequencing. The four selected genomes represent diverse and genomically unexplored sections of the *Aspergillus* genus: *A. campestris* is the first member of section *Candidi* to be sequenced, and likewise *A. steynii* is the first member of section *Circumdati* to be sequenced. *A. ochraceoroseus*, the first member of section *Ochraceorosei*, has recently been draft genome sequenced (3) and is available only in a large number of scaffolds. Here we also present a greatly improved assembly that may serve as a reference genome for this section. Furthermore, we have added a highly interesting member of section *Fumigati*, *A. novofumigatus*, which has a diverse secondary metabolite (SM) profile (4), as well as potentially being an opportunistic pathogen with close relation to the medically very important *A. fumigatus* (5). In addition, two strains from the *Candidi* section were Illumina sequenced to elucidate the chlorflavonin biosynthesis.

Significance

The genus of *Aspergillus* holds fungi relevant to plant and human pathology, food biotechnology, enzyme production, model organisms, and a selection of extremophiles. Here we present six whole-genome sequences that represent unexplored branches of the *Aspergillus* genus. The comparison of these genomes with previous genomes, coupled with extensive chemical analysis, has allowed us to identify genes for toxins, antibiotics, and anticancer compounds, as well as show that *Aspergillus novofumigatus* is potentially as pathogenic as *Aspergillus fumigatus*, and has an even more diverse set of secreted bioactive compounds. The findings are of interest to industrial biotechnology and basic research, as well as medical and clinical research.

Author contributions: I.K., T.C.V., J.C.F., E.K.L., M.E.K., K.B., B.A.S., J.K.M., U.H.M., T.O.L., I.V.G., S.E.B., and M.R.A. designed research; I.K., T.C.V., E.K.L., M.E.K., C.Y.N., C.D., and J.C. performed research; I.K., A.S., B.A.S., J.K.M., and T.O.L. contributed new reagents/analytic tools; I.K., T.C.V., J.C.F., J.L.N., S.T., A.K., P.B., Y.M., S.M., A.C., A.L., A.S., K.L., S.H., T.O.L., S.E.B., and M.R.A. analyzed data; and I.K., T.C.V., J.C.F., J.L.N., S.T., P.B., Y.M., S.M., M.E.K., U.H.M., T.O.L., I.V.G., and M.R.A. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: All the sequencing data are available at the JGI Genome Portal (genome.jgi.doe.gov). *A. campestris* (accession no. MSFM00000000) genome.jgi.doe.gov/Aspcam1/Aspcam1.home.html *A. novofumigatus* (accession no. MSZS00000000) genome.jgi.doe.gov/Aspnov1/Aspnov1.home.html *A. ochraceoroseus* (accession no. MSFN00000000) genome.jgi.doe.gov/Aspoch1/Aspoch1.home.html *A. steynii* (accession no. MSFO00000000) genome.jgi.doe.gov/Aspste1/Aspste1.home.html A. candidus (accession no. PKFS00000000) genome.jgi.doe.gov/Aspcand1/Aspcand1.home.html *A. taichungensis* (accession no. PKFW00000000) genome.jgi.doe.gov/Asptaic1/Asptaic1.home.html.

[1]To whom correspondence should be addressed. Email: mr@bio.dtu.dk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1715954115/-/DCSupplemental.

The four PacBio sequenced species explored in this study can act as a reference strain in their respective phylogenetic sections. The species may also be used to assess the natural variation within the *Aspergillus* genus via analysis of species-specific genes in comparison with other genome-sequenced species. Accordingly, we have compared our sequenced genomes with nine published *Aspergillus* reference genomes (from sections *Nidulantes*, *Nigri*, *Fumigati*, *Flavi*, *Clavati*, and *Terrei*) to serve as a compilation of reference strains for the genus.

In addition, the biosynthetic potential of these species is of interest. Filamentous fungi produce a diverse range of SMs, including bioactive compounds such as pharmaceuticals and toxins (6). SMs are not required for growth, but provide important benefits in the growth environment (7). Members of the *Aspergillus* genus are known to produce a wide variety of SMs with industrial, agricultural, medical, and economic importance (7, 8). The biosynthetic genes of SMs are located in clusters setting the stage for common gene regulation (9, 10). Clusters often span tens of kilobases (kbs) (11) and usually contain a gene or genes coding for one or more synthases (backbone enzyme) that define the product class of the cluster [i.e., polyketide synthases (PKS), nonribosomal peptide synthetases, and prenyltransferases or terpene cyclases (12)], in addition to tailoring enzymes such as transferases, hydroxylases, and regulatory proteins and transporters (11, 12).

With the increasing number of whole-genome sequences, the opportunity of performing analysis based on comparative genomics arises, which can give important insights and knowledge. With a focus on investigating bioactive and toxic compounds, we have here identified biosynthetic gene clusters responsible for interesting compounds from each of the PacBio sequences by combining genome analysis with knowledge of biochemical pathways and compound structure. We have identified candidates for the ochrindol cluster in *A. steynii*, and the chlorflavonin cluster in *A. campestris*.

A. novofumigatus was investigated on a genetic level, focusing on SMs. The secondary metabolic potential has been investigated, and biosynthetic gene clusters for three compounds (novofumigatonin, *epi*-aszonalenin, and *ent*-cycloechi) have been identified. Furthermore, the genomic differences and similarities of the closely related species *A. novofumigatus* and the pathogen *A. fumigatus* have been investigated, focusing on SMs, allergens, and virulence factors, and thereby addressing the potential pathogenicity of *A. novofumigatus* and how closely related the two morphologically similar species are.

In addition, the evolution of the aflatoxin (a highly carcinogenic compound) gene cluster from *A. ochraceoroseus* was investigated. The biosynthetic gene cluster was identified and studied earlier in several species, including *A. flavus*, *A. parasiticus*, and *A. ochraceoroseus* (13, 14). It has been seen that the synteny of the clusters are quite varying and that *A. ochraceoroseus* is missing some essential genes (*aflQ* and *aflP*) in the biosynthesis of aflatoxin known from *A. flavus* (14). With whole-genome sequences at hand, we have addressed some of these questions concerning the evolution of this biosynthetic gene cluster.

Results and Discussion

Genome Statistics. The genomes of *A. campestris*, *A. novofumigatus*, *A. ochraceoroseus*, and *A. steynii* were sequenced using PacBio RS, whereas *A. taichungensis* and *A. candidus* were sequenced using Illumina (see *SI Appendix* for details). Annotation of the genomes was completed using the JGI Annotation Pipeline (15). Table 1 lists genome sequence statistics for each of the six species. The four PacBio sequenced genomes have a relatively low number of scaffolds and do not contain internal gaps. For that reason, they are highly useful as references for comparative genomics, as well as for studies of the individual genomes. Of the sequenced genomes, *A. steynii* has the largest genome size and is comparable with that of *A. oryzae* (16). The genome of *A. steynii* is ∼27% larger than *A. ochraceoroseus*, which has the smallest genome in this set and has a genome size comparable with *A. clavatus* (17). The difference in genome size also reflects the numbers of predicted genes in the two species, which range from 13,211 to 8,924, respectively.

Investigation of DNA Methylation. Because the four *Aspergillus* genomes (*A. steynii*, *A. campestris*, *A. novofumigatus*, *A. ochraceoroseus*) have been sequenced using PacBio, it is possible to investigate the presence of N6-methyldeoxyadenine (6mA) (18). Previous attempts at validation of such low abundance of 6mA have proven challenging, making it difficult to conclude whether 6mA is present in these fungi and, if so, to discriminate between real 6mA sites and false-positives (18). The presence of 6mA was therefore explored across the four *Aspergillus* genomes (Table 2). Consistent with previous reports (18) of low levels of 6mA in the Dikarya, we detect very little 6mA in the Aspergilli, ranging from 0.012 (*A. steynii*) to 0.038 (*A. campestris*) percent adenines methylated compared with early-diverging fungi, in which up to 2.8% of all adenines were methylated (Table 2) (18). Furthermore, only a handful of 6mA sites were at ApT dinucleotides, and none was found symmetrically at ApTs, both of which are characteristic features of 6mA modification in early-diverging fungi (18). The results therefore suggest an absence or very low occurrence of 6mA methylation in Aspergilli.

Whole-Genome Phylogeny Confirms Species Found in Separate Clades. To provide an overview of the relationships among the sequenced species in the *Aspergillus* genus, we constructed a phylogenetic tree of the four PacBio sequenced species and the 11 reference strains, including *Penicillium chrysogenum* and *Neurospora crassa* as outgroups (Fig. 1).

The constructed phylogenetic tree supports the results described earlier by Peterson (21), where a tree was constructed based on DNA sequences of four loci. *A. campestris* most closely resembles *A. terreus* of the reference genomes, whereas *A. steynii* relates closest to *A. flavus* and *A. oryzae*. Members of the *Fumigati*

Table 1. Overview of sequencing and annotation data for the four investigated PacBio-sequenced species, plus two additional Illumina-sequenced species

	A. campestris	*A. novofumigatus*	*A. ochraceoroseus*	*A. steynii*	*A. candidus*	*A. taichungensis*
Genome size, Mbp	28.3	32.4	27.7	37.8	27.3	27.12
Number of proteins	9,764	11,549	8,924	13,211	9,641	9,692
Number of scaffolds	62	62	34	37	268	310
Number of scaffolds ≥2 kbp	56	62	32	36	168	283
Scaffold N50	6	4	4	4	23	47
Scaffold L50	1,703,432	3,768,347	2,489,623	3,921,250	391,998	207,690
Fraction of GC, %	51.2	49.1	44.2	49.1	51.8	51.44
Coverage of gaps, %	0	0	0	0	0.0298	0.0155
Coverage of InterPro, %	68	67	67	66	75	25

Table 2. Overview of the methylation pattern of *A. campestris*, *A. novofumigatus*, *A. ochraceoroseus*, and *A. steynii*

Lineage	Percentage adenines methylated	Total number of sites	Percentage modifications at ApT sites
A. steynii	0.012	6,753	0.054
A. campestris	0.038	9,156	0.041
A. novofumigatus	0.03	7,917	0.058
A. ochraceoroseus	0.021	7,355	0.027

section are in a single clade (marked in blue on Fig. 1), with *A. clavatus* as a close relative. *A. ochraceoroseus* is placed next to *A. nidulans*, and both belong to the subgenus Nidulantes. All the species belonging to subgenus Circumdati (*A. niger*, *A. oryzae*, *A. flavus*, *A. steynii*, *A. terreus*, and *A. campestris*) are also placed in one clade. The tree further confirms that the three species *A. ochraceoroseus*, *A. steynii*, and *A. campestris* indeed represent distinct branches in the *Aspergillus* phylogram (22).

Unique Genes in the Genomes Often Encode Regulatory Proteins and Enzymes Involved in Secondary Metabolism. We have identified and investigated species-specific genes for the four newly sequenced species to examine the diversity within the *Aspergillus* genus. Genes that are unique to a species or a small group of species may be associated with phenotypic traits and adaptation of these species to specific environments. We define species-specific genes as those without any orthologs in other sequenced genomes. This definition makes the set of species-specific genes dependent on the strains included in the analysis. As more genomes are included, especially genomes from closely related species or strains, fewer species-specific genes will be identified. The species-specific genes for each genome were identified using a set consisting of the four PacBio sequenced genomes and 11 reference genomes (*SI Appendix*, Table S1). Two closely related strains will share most of their genes, and they will as such not be unique to the individual species. The unique genes are not expected to encode any key functions in the cell, as they are found in only one organism; instead, these genes might be involved in environmental adaptation and/or speciation. The strains have 22%, 15%, 21%, and 27% unique genes for *A. campestris*, *A. novofumigatus*, *A. ochraceoroseus*, and *A. steynii*, respectively, indicating the vast diversity found within the *Aspergillus* genus. Approximately one third of the species-specific genes could be associated with an InterPro sequence domain (*SI Appendix*, Table S1) (23), suggesting that these genes are not false annotations.

Comparative Analysis of the Genomes of *A. novofumigatus* and *A. fumigatus*. *A. novofumigatus* and *A. fumigatus* are considered to be two closely related species, and *A. novofumigatus* has only been regarded a separate species since 2005 (24). The homology between *A. novofumigatus* and *A. fumigatus* has been investigated based on the number of *A. novofumigatus* proteins with BLASTP hits (≥50% identity ≥130% coverage of query plus hit) in *A. fumigatus*. Based on this, 8,385 of *A. novofumigatus* proteins have homologous proteins in *A. fumigatus*, corresponding to 73%. The synteny between the two species was also examined using NUCmer (Nucleotide Mummer) from the MUMmer 3.0 package to map *A. novofumigtus* genome to the reference genome of *A. fumigatus* (25–27). Based on these alignments, 23.1 Mbp of the *A. novofumigatus* genome can be mapped to *A. fumigatus*, corresponding to 71% of the *A. novofumigatus* genome. The maximum block size is 75 kbp, and the mean block size is 4.6 kbp.

To explore this difference genetically and functionally, we have explored the similarities and differences between these two species with a focus on allergens, genes involved in virulence, and production of SMs.

Secondary Metabolite Profile of *A. novofumigatus* Compared with *A. fumigatus*. The extrolite production in *A. fumigatus* has been extensively studied, and an abundance of SMs have been identified (4). *A. novofumigatus* is also known to have a versatile secondary metabolism; however, there is very little overlap of extrolite production between the two closely related species, making it very interesting to compare their genetic potential for producing SMs (4).

To investigate what type of SM gene clusters *A. fumigatus* and *A. novofumigatus* have in common, the SM gene clusters were predicted for each genome, using an implementation of SMURF (28). An overview of the predicted clusters and homologs in *A. novofumigatus* and *A. fumigatus* is presented in Fig. 2 *A* and *B*, respectively. Of the 34 predicted clusters in *A. fumigatus* and 56 predicted clusters in *A. novofumigatus*, 24 appear to be shared among the two species, based on bidirectional BLAST hits of the synthase (Fig. 2*C*). Of the 11 elucidated clusters from *A. fumigatus* [based on MIBiG (29)], homologs of seven (Gliotoxin, hexadehydro-astechrome, pseurotin A, fumagillin, endocrocin, helvolic acid, and trypacidin) can be found in *A. novofumigatus* (based on homology of the synthase). Several of these SMs are known to be involved in the virulence of *A. fumigatus* and are examined in more detail in the *Comparative Genomics of Genes Encoding Allergens, Virulence, and Pathogenicity Factors*.

The prediction of SM gene clusters also revealed considerable differences between the two closely related species. First, as seen in Fig. 2*A*, *A. novofumigatus* has 17 predicted clusters with no orthologs in any of the reference species. This is in contrast to *A. fumigatus*, which has only three clusters without orthologs in the reference species (Fig. 2*B*). Second, *A. novofumigatus* has 56 predicted SM clusters, whereas *A. fumigatus* has only 34 (Fig. 2*D*). Third, *A. novofumigatus* also has more different types of clusters. An overview of the cluster types and the number of clusters found in the two species can be seen in Fig. 2*D*. The diversity of SM gene clusters supports the identification of these two organisms as separate species.

SMs can present a competitive advantage in the battle for resources, but if the environment is stable, there is no need for a large arsenal of different metabolites. Thus, the large difference in SM potential between *A. fumigatus* and *A. novofumigatus* might be a reflection of the difference in natural environment and the competition in these environments, indicating that *A. novofumigatus* normally exists in a highly competitive environment and has a need for a larger repertoire of SMs. These results do not suggest in which conditions the metabolites are produced. Perhaps the differences are influenced by that fact that *A. fumigatus* Af293 is from a clinical isolate, whereas *A. novofumigatus* have been isolated from chamise chaparral soil after a bush fire in Southern California (2, 24). Indeed, earlier analyses have shown that clinical isolates produce fewer exometabolites and sporulate less (4, 30, 31).

It is clear that the sequence of *A. novofumigatus* represents a significant number of unknown gene clusters, and thereby possibly interesting bioactive compounds. To start explore this treasure chest and to illustrate our approach of linking metabolites to their respective gene clusters, we have here identified four highly interesting compounds (novofumigatonin, *ent*-cycloechinulin, and *epi*-aszonalenin A and C) by liquid chromatography–mass spectrometry analysis (*SI Appendix*, Fig. S1), and we have identified the biosynthetic gene clusters by using comparative genomics. Our analysis targeted these four model compounds, as they represent major metabolites produced by *A. novofumigatus* and because we have them as pure standards in our in-house collection of fungal metabolites (32).

Novofumigatonin is chemically a very complex compound containing an orthoester and is at present only known to be produced by *A. novofumigatus* (33). It has been suggested that novofumigatonin

is a meroterpenoid produced from the aromatic polyketide, 3,5-dimethylorsellinic acid, as an initial precursor. Terretonin is another fungal meroterpenoid derived from 3,5-dimethylorsellinic acid, and thus we hypothesized that the early-stage biosynthetic route for novofumigatonin would follow the same pathway as that for terretonin, found in *A. terreus* (34).

The terretonin biosynthetic genes in *A. terreus* were used to find homologs in *A. novofumigatus*, using BLASTP. Exactly one candidate cluster was identified (Fig. 3*C*), containing orthologs to all of the genes from the early stages of the terretonin pathway, with one exception: a terpene cyclase (*trt1*). However, we found this just upstream of the identified predicted cluster, and it was thus included in the putative cluster and in Fig. 3*C*.

Novofumigatonin is closely related to fumigatonin, which has been reported to be produced by *A. fumigatus* (35). Interestingly, no similar cluster could be found in *A. fumigatus*. The most similar cluster in *A. fumigatus* (which is not a very strong hit: three proteins with amino acid identity <50%) is an already known cluster responsible for production of another meroterpenoid, pyripyropene A (36). This is very puzzling, and one could speculate that the report of fumigatonin might have been obtained from a misidentified isolate supposed to be *A. fumigatus* (35), but in reality an *A. novofumigatus* strain, which was only recently described as a separate species (24).

A putative cluster (scaffold 3, 33,642–53,133 bp) for *ent*-cycloechinulin (*SI Appendix*, Fig. S2) in *A. novofumigatus* was identified using the fumitremorgins (*ftm*) cluster from *A. fumigatus* as the starting point. Fumitremorgins are similar to *ent*-cycloechinulin (*SI Appendix*, Fig. S2), but with some important differences. *Ent*-cycloechinulin uses alanine instead of proline as a starter unit. Furthermore, the following prenylation occurs in a reverse manner. The genes responsible for these steps therefore have a low identity (≤35%). The following hydroxylation and *O*-methylation, however, are more similar, which is also reflected in the identity for the genes (≥65%). Even though the identity for the identified genes is low, they still represent the best hits in *A. novofumigatus*, supporting that this is the best candidate cluster.

In a similar fashion, a putative cluster for *epi*-aszonalenin A and C in *A. novofumigatus* was identified using a similar acetylaszonalenin cluster (*SI Appendix*, Fig. S2 for chemical structure) from *A. terreus*. Again, the acetylaszonalenin cluster proteins were used for a comparative genomics search in *A. novofumigatus*. This way, a very similar cluster was identified (scaffold 3, 1,663,448–1,719,848 bp) as a candidate for *epi*-aszonalenin A and C biosynthesis.

Comparative Genomics of Genes Encoding Allergens, Virulence, and Pathogenicity Factors. *A. fumigatus* is known to be a common opportunistic human pathogen (37), whereas *A. novofumigatus* has only been reported as pathogenic in one instance (5). This difference in known pathogenicity offers the opportunity to identify and compare allergens and genes involved in the pathogenicity based on orthology, and to gather insights into the potential harmfulness of *A. novofumigatus*.

A list of currently accepted allergenic proteins from the well-studied *A. fumigatus* can be extracted from the Allergome database (www.allergome.org). The sequences of the allergenic proteins from this list were compared with the annotated *A. novofumigatus* protein list using BLAST+, with parameters set to report full-length

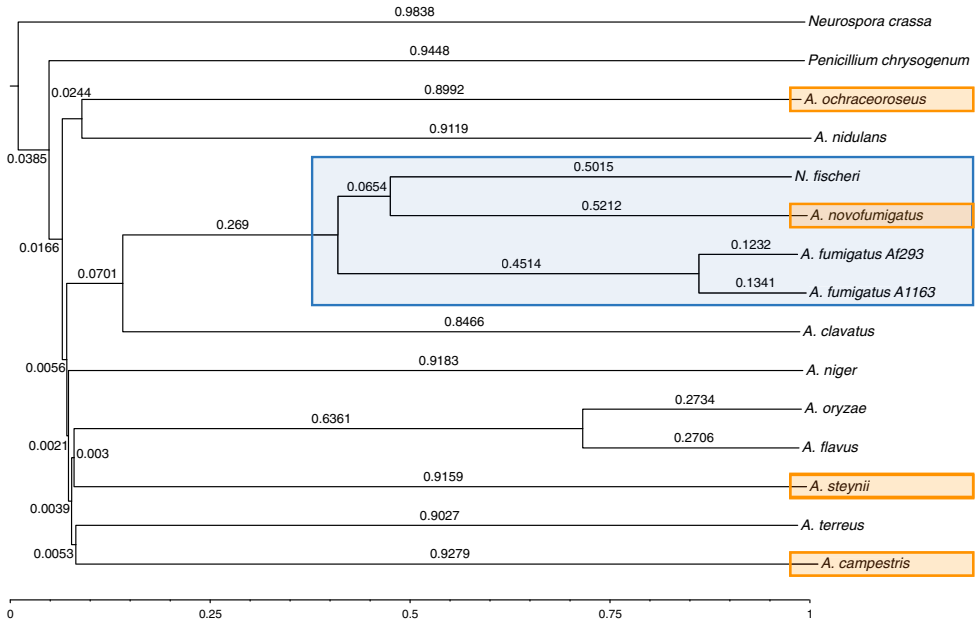


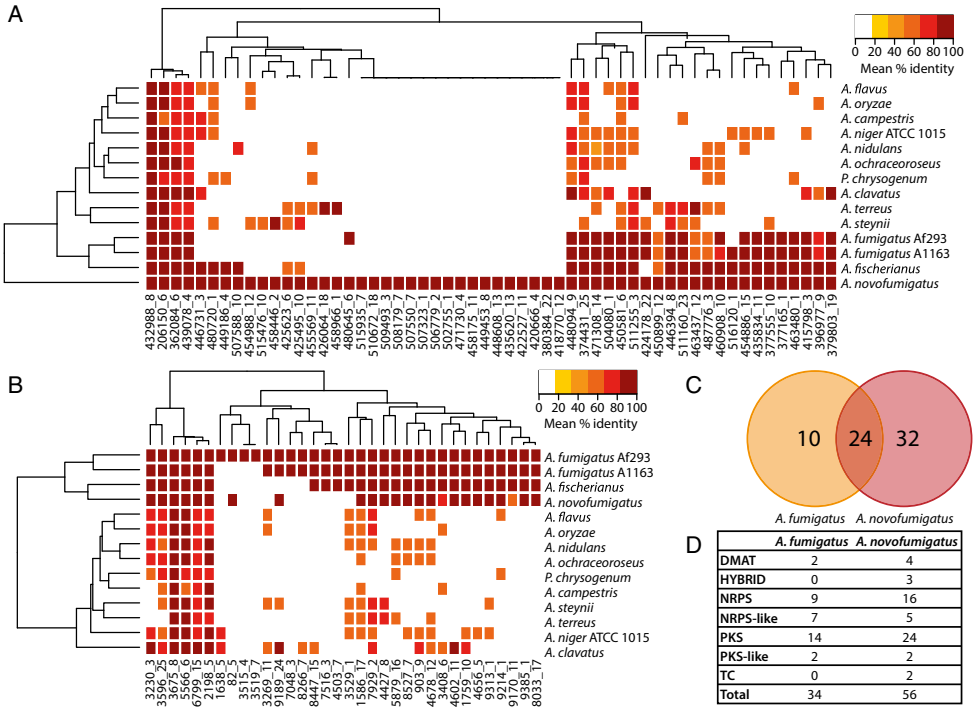
Fig. 1. Phylogenetic tree of 15 strains constructed from a composition vector approach of whole-proteome sequences using CVTree3 with a K′mer size of eight (19, 20). The time scale has been scaled to the root, thereby making the branch lengths relative to the distance between the root and the species. This time will therefore depend on what species are included in the set. The four PacBio-sequenced species are marked with orange, and the *Fumigati* section is marked with blue.

sequence matches. Results shown in *SI Appendix*, Table S2 indicate that all *A. fumigatus* allergen proteins are represented in the *A. novofumigatus* genome. Of a total of 41 proteins, 34 proteins showed >90% identity, four showed 85–90% identity, and three showed 50–80% identity. As proteins with >50% identity are likely to cross-react to IgE (38), these results strongly indicate that *A. novofumigatus* possesses a strong allergen repertoire that will at least cross-react strongly with IgE to *A. fumigatus* and is likely to be able to provoke an immune response in the same manner as *A. fumigatus*. It is not possible to rule out the possibility that *A. novofumigatus* could be a more virulent pathogen or allergenic sensitizer than *A. fumigatus*.

A set of 35 potential virulence genes was assembled from recent literature, as well as genes responsible for biosynthesis of the SMs melanin, fumagillin, fumitremorgins, gliotoxin, and helvolin, which are reported to play a direct role in virulence (4, 39, 40). The results are shown in *SI Appendix*, Table S3. The majority of the potential virulence genes are shared between *A. fumigatus* and *A. novofumigatus* with high similarity (>85% identity); only *arp2* and *gel2* had identity just below 50%. The fumitremorgins cluster consists of nine genes, six of which have identity <50%, including the synthase indicating that *A. novofumigatus* is unable to produce fumitremorgins. The two SM gene clusters for gliotoxin and fumagillin in *A. fumigatus* both have highly similar matches in *A. novofumigatus*. The cluster for helvolic acid has three genes of nine with low BLASTP identity of 42–48%. However, *A. novofumigatus* has been reported to produce helvolic acid, indicating that a high amino acid similarity of these genes is not required (4).

It is likely that different combinations of virulence factors among the species affect pathogenicity (31). It has been suggested that species unable to produce some metabolites may be able to produce proxy-exometabolites that can serve the same function. This could indicate that species producing many different kinds of exometabolites are potentially pathogenic (4).

A. novofumigatus possesses the full range of allergen proteins expressed by *A. fumigatus*, in addition to the majority of virulence factors including several SMs. Furthermore, *A. novofumigatus* has an extensive potential for SM production with 56 predicted gene clusters compared with 34 for *A. fumigatus*. Together, these results indicate that *A. novofumigatus* has a considerable potential to be pathogenic. The observation of only a single instance of invasive infection by *A. novofumigatus* (5) may result from the recent development of methods to identify this species, which has previously not been distinguishable from *A. fumigatus*. It has been found that ∼4–5% of *A. fumigatus* isolated from patients later turned out to be closely related species (41). Thus, the true pathogenic potential of *A. novofumigatus* might be underestimated. Similarly, allergen sensitization to *A. novofumigatus* is not currently



	A. fumigatus	*A. novofumigatus*
DMAT	2	4
HYBRID	0	3
NRPS	9	16
NRPS-like	7	5
PKS	14	24
PKS-like	2	2
TC	0	2
Total	34	56

Fig. 2. (*A*) Overview of the SM gene clusters predicted in *A. novofumigatus* and their homologs in the reference species. (*B*) Overview of the SM gene clusters predicted in *A. fumigatus* and their homologs in the reference species. (*C*) A Venn diagram of the *A. fumigatus* and *A. novofumigatus* SM gene clusters. (*D*) The number and different types of SM gene clusters predicted in *A. fumigatus* and *A. novofumigatus*. DMATS, dimethylallyl tryptophan synthase; NRPS, nonribosomal peptide synthetase; PKS, polyketide synthase; TC, terpene cyclase.

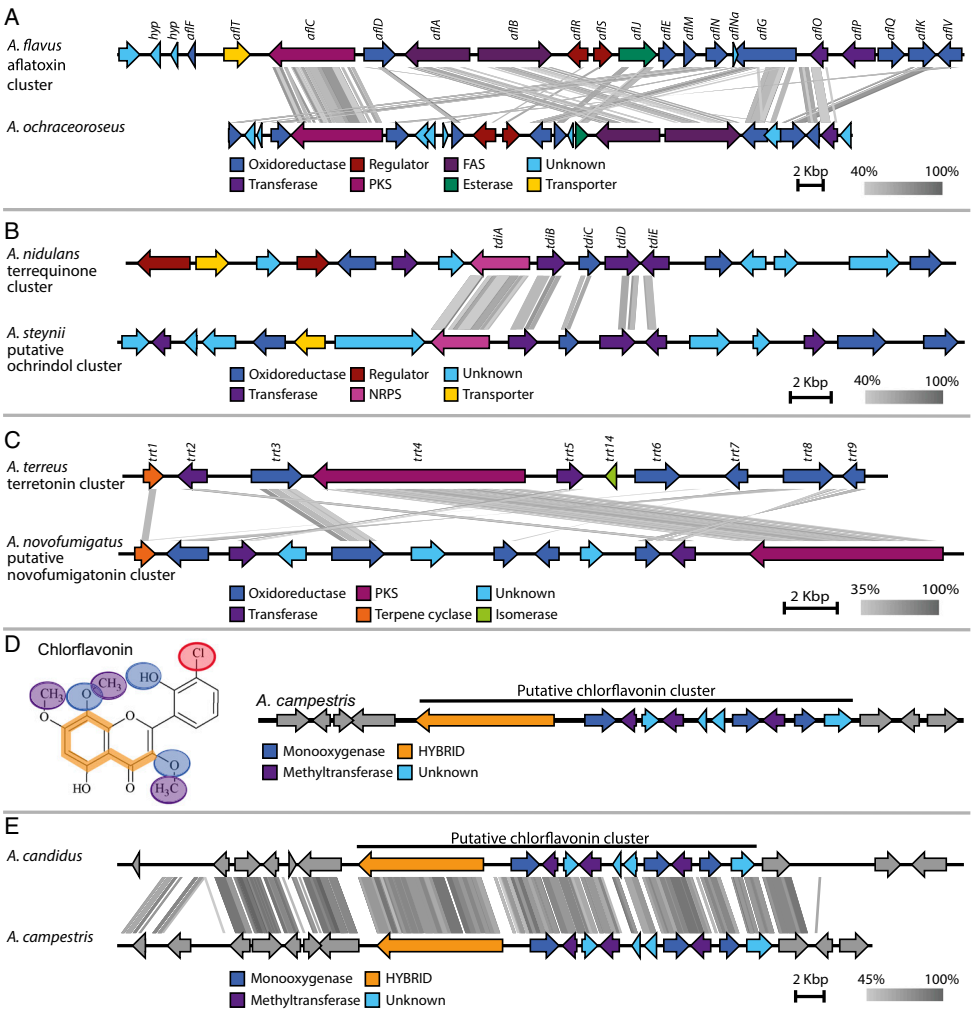


Fig. 3. Synteny plots of investigated clusters made using Easyfig tBLASTx. (*A*) The synteny of the predicted aflatoxin cluster in *A. flavus* NRRL3357 and the identified candidate aflatoxin cluster in *A. ochraceoroseus* (scaffold 2, 4,201,774–4,251,209 bp). (*B*) Synteny plot of the candidate cluster for ochrindol in *A. steynii* (scaffold 7, 2,783,445–2,824,507 bp) and terrequinone cluster in *A. nidulans*. The terrequinone cluster consists of a single-module nonribosomal peptide synthetase (tdiA), a prenyltransferase (tdiB), an oxidoreductase (tdiC), an aminotransferase (tdiD), and a gene of unknown function similar to a methyl transferase (tdiE). (*C*) Synteny plot of the known meroterpenoid cluster of terretonin in *A. terreus* and the candidate cluster of novofumigatonin in *A. novofumigatus* (scaffold 14, 103,246–136,450 bp). (*D*) The chemical structure of chlorflavonin and the candidate cluster for chlorflavonin in *A. campestris* (scaffold 1, 576,100–603,958 bp). The hydroxylation has been highlighted with blue, the O-methylation has been highlighted in purple, and the PKS backbone has been highlighted in orange. (*E*) Synteny plot of putative chlorflavonin clusters in *A. candidus* and *A. campestris*.

tested, and this species may also have potential to contribute to the burden of fungal allergy.

Investigation and Evolution of the Aflatoxin Gene Cluster in *A. ochraceoroseus*. It is well known that *A. ochraceoroseus* can produce aflatoxin, and the biosynthetic cluster has been identified (14). Also, it has been noted that the aflatoxin gene cluster in *A. ochraceoroseus* is missing homologs to the *aflP* and *aflQ* gene involved in the conversion of sterigmatocystin (ST) to aflatoxin.

Here we have compared the aflatoxin gene cluster from the whole-genome-sequenced *A. flavus* NRRL3357 with *A. ochraceoroseus*. The

clusters were identified in both species by using the aflatoxin genes identified in *A. flavus* AF70 (AY510453) (42).

Comparing the two clusters from *A. flavus* NRRL3357 with *A. ochraceoroseus*, it is evident that the synteny is characterized by gene shuffling (Fig. 3*A*). The identified cluster in *A. ochraceoroseus* is more similar to the ST gene cluster known from *A. nidulans* in the organization of genes, which was also the result of previous findings (14). This is evolutionary very interesting, as the clusters producing the same compound are quite different in their synteny, suggesting cluster dynamics or distant evolutionary origin.

As found by Cary et al. (14) it was seen that the *A. flavus aflP* and *aflQ* genes are missing in the *A. ochraceoroseus* aflatoxin cluster. These genes are important for the biosynthesis of aflatoxin. The whole-genome sequence was searched for orthologs to the *aflP* and *aflQ* genes from *A. flavus*, using BLASTP. The best hit for *aflQ* was JGI protein 547596, with identity 56.3% and coverage of 95.3%. The best hits for *aflP* were JGI proteins 430163, 506769, and 427152, with identity ranging from 40.5% to 36.6% and coverage between 31.4% and 37.7%. All the potential genes are located on a different scaffold than the aflatoxin cluster. The genes identified here are possible candidates for the *A. ochraceoroseus* version of the *aflP* and *aflQ* genes. With this information, it is not possible to determine exactly which genes are responsible for the conversion from ST to aflatoxin, but based on homology, these are the best candidates. Another possibility could be that the *aflP* and *aflQ* genes in *A. ochraceoroseus* have arisen via convergent evolution, and would thus not be found via homology analysis.

In summary, the identified aflatoxin gene cluster in this *A. ochraceoroseus* genome shows that *A. ochraceoroseus* and *A. flavus* most likely represent various stages of the aflatoxin cluster evolution. However, to get the full picture and truly understand the evolution of the clusters, more aflatoxin and sterigmatocystin producers need to be sequenced to be able to make bigger comparisons and get a better idea of where and when the different variations were created.

Identifying the Ochrindol Cluster in *A. steynii*. Ochrindoles are prenylated bisindolyl benzoid/quinone metabolites (*SI Appendix*, Fig. S2) that have shown anti-insectant properties (43), one reason that *A. steynii* is an interesting species. Ochrindol is produced by *A. steynii*, and the chemical structure is known, but the biosynthetic pathway is unknown (44). However, the biosynthesis of a similar compound, terrequinone (*SI Appendix*, Fig. S2), produced by *A. nidulans*, is known, and so are the five biosynthetic genes *tdiA–tdiE* (45). It has been shown that ochrindol D is produced as an intermediate during biosynthesis of terrequinone. We therefore hypothesize that the genes for the biosynthesis would be partly similar, and could thus be used to identify the ochrindol cluster.

First, the five *tdi* genes were identified in the *A. nidulans* genome in a predicted cluster consisting of 17 genes. Significantly, five genes similar to *A. nidulans tdiA–tdiE* were identified in a predicted cluster of 17 genes, with the synteny of the *tdiA–tdiE* orthologs conserved (Fig. 3*B* and *SI Appendix*, Table S4). However, none of the genes next to the five *tdi* genes showed any homology or synteny, suggesting that the size of the cluster is overpredicted, at least in *A. nidulans*. In *A. steynii*, some of the extra genes could be involved in ochrindol production.

Identifying the Chlorflavonin Cluster in *A. campestris*. Chlorflavonin was the first fully characterized flavone with fungal origin, and it is also the first naturally occurring flavone discovered to be chlorinated. It has been shown to have antifungal properties against specific species (46). The chemical structure of chlorflavonin (*SI Appendix*, Fig. S2) is known, and a biosynthetic pathway has been proposed, but no genes associated with the biosynthesis have been identified (47). With the whole-genome sequence for *A. campestris* at hand, we started exploring the genetic potential to identify the biosynthetic gene cluster responsible for producing this interesting compound.

Initially, looking at the chemical structure of this fungal flavonoid, an obvious idea for the biosynthesis would be that the backbone structure is created by a type III PKS, as the compound is so similar to plant flavonoids produced by type III PKS (48, 49). However, no type III PKS were found in *A. campestris*, suggesting a fungal-specific mode of biosynthesis. Next, investigating the chemical structure and proposed general biosynthesis for chlorflavonin (47), it could be deduced that the cluster must contain at least a PKS/hybrid backbone, three monooxygenases, three methyltransferases, and a chlorinating enzyme. Only one cluster met the requirements of three monooxygenases and three methyltransferases (Fig. 3*D*). The only concern with this candidate cluster is the lack of the essential chlorinating enzyme (*SI Appendix*, Table S5, Part 3).

First, sequences of known chlorinating enzymes (*SI Appendix*, Table S5, Part 1) were used to search for similar proteins in *A. campestris*, using BLASTP, but no genes were found (51–54). Second, relevant possible chlorinating InterPro domains were identified and found in four genes (*SI Appendix*, Table S5, Part 2), although it was not possible to pinpoint the best candidate of the chlorinating enzyme with these methods. However, the identified cluster is currently the best candidate cluster for chlorflavonin in *A. campestris*. Verification of this by knock-out experiments or heterologous expression could verify the candidate clusters as being responsible for the production of chlorflavonin, but this organism is not currently genetically engineerable, and the gene cluster is too large to transfer.

We therefore set out to support our prediction by sequencing and comparing genomes of several closely related species from section Candidi. *A. candidus* is a known chlorflavonin producer, whereas *A. taichungensis* is not (50). These species were therefore whole-genome sequenced to compare the pattern of the producers with the predicted clusters. *A. campestris*, *A. candidus*, and *A. taichungensis* each have 48, 45, and 43 predicted clusters. Based on the backbone, *A. campestris* and *A. candidus* share 35 clusters and *A. campestris* and *A. taichungensis* share 31 (BLASTP ≥50% identity and ≥130% hit+query coverage).

Comparing the genes found in the putative chlorflavonin cluster in *A. campestris* with the whole-genome sequences of *A. candidus* and *A. taichungensis*, *A. candidus* was homologous to genes in the putative chlorflavonin cluster (Fig. 3*E*). Moreover, this cluster is also the only cluster in *A. candidus* that has three methyltransferases and three monooxygenases. *A. taichungensis*, in contrast, does not have any significant hits of the predicted biosynthesis genes, as would be expected.

In addition, the chlorinating potential was investigated in these species. As with *A. campestris*, there were no BLASTP hits in *A. candidus* and *A. taichungensis* from the known chlorinating proteins (*SI Appendix*, Table S5, Part 1) (51–54).

Also, the possible chlorinating InterPro domains were investigated in the genomes of *A. candidus* and *A. taichungensis*. The number of hits were similar; however, *A. campestris* had one more hit for IPR001568, and both *A. campestirs* and *A. candidus* had one more hit for IPR008775, but none of the hits is found in SMURF-predicted clusters (*SI Appendix*, Table S5, Part 2).

These investigations further support that the identified cluster in *A. campestris* is the best candidate for chlorflavonin biosynthesis.

Conclusion

In this study, high-quality PacBio genome sequence data were generated for four *Aspergillus* species (*A. campestris*, *A. novofumigatus*, *A. ochraceoroseus*, and *A. steynii*) and investigated using comparative genomics. Furthermore, we have prepared draft genome sequences for two additional species: *A. taichungensis* and *A. candidus*.

These six species are diverse and represent various sections of the *Aspergillus* genus, and thereby provide insight into the genomic and biochemical diversity and potential of the genus.

The four PacBio sequenced species have been compared with a group of already whole-genome-sequenced *Aspergillus* species to determine the level of genetic diversity. A phylogram was constructed on the basis of the whole-genome proteomes, and the resulting tree supports the taxonomy of the genus and fits with a phylogenetic tree constructed by Peterson SW (21) and Kocsubé S, et al. (22), based on four loci or nine loci (21, 22). The tree confirms that *A. campestris*, *A. ochraceoroseus*, and *A. steynii* indeed represent sections of the *Aspergillus* genus, which have not been genome sequenced before. Analysis of the genomes show that these genomes represent a large number of species-specific genes, particularly within secondary metabolism.

Investigation of the presence of N6-methyldeoxyadenine of the four presented species shows very low levels of 6mA. Moreover, no 6mA sites were found symmetrically at ApTs, which has been found to be a characteristic feature of 6mA modification in early-diverging fungi (18), thus confirming previous suggestions that 6mA methylation is not significant in Aspergilli.

A. novofumigatus has been compared with a close relative, the pathogenic species *A. fumigatus*, to better understand the mechanism of pathogenicity and virulence. The predicted SM gene clusters were found to be very different for the two close relatives, with *A. novofumigatus* containing 65% more clusters than *A. fumigatus*.

All allergens known from *A. fumigatus* are also present in *A. novofumigatus*, and the majority of the virulence factors are shared between the two species. The major difference is that *A. novofumigatus* lacks the fumitremorgin cluster. However, it has been suggested that proxy-exometabolites may serve the same function and *A. novofumigatus* has an extensive arsenal of additional SM gene clusters. It is thus highly likely that *A. novofumigatus* is a highly capable pathogen.

Furthermore, we have, with multiple examples, demonstrated that it is possible to identify the respective gene cluster using whole-genome sequences if one has a well-established structure of a SM and biological and chemical insights to the pathway. This way we have reidentified the aflatoxin gene cluster in *A. ochraceoroseus*; the *epi*-aszonalenins, novofumigatonin, and *ent*-cycloechinulin gene clusters in *A. novofumigatus*; the ochrindol cluster in *A. steynii*; and finally, the chlorflavonin cluster in *A. campestris*, backed by additional info from sequencing the *A. taichungensis* and *A. candidus* genomes.

In summary, the six genome sequences presented in this study illustrate the large diversity found in the *Aspergillus* genus and highlight the potential for discovery of structurally diverse SMs. As our project of sequencing +300 species progresses along with other fungal genome sequencing projects (e.g., the 1K fungal genomes project 1000.fungalgenomes.org/home/), the potential for applying comparative genomics to get evolutionary insights and discover interesting SMs will only increase.

Materials and Methods

The materials include a list of sequenced strains. Methods include strain cultivation; genome sequencing, assembly, and annotation; DNA-methylation analysis; details for comparative genomics analysis; phylogeny; and chemical analysis of secondary metabolism. Details for all methods are found in *SI Appendix*, *SI Text*. In particular, we provide a detailed protocol for efficient, reproducible, and scalable DNA and RNA extraction from fungi.

ACKNOWLEDGMENTS. M.R.A. and T.C.V. gratefully acknowledge funding from the Villum Foundation, Grant VKR023437. Genome sequencing was kindly supported by Joint BioEnergy Institute and Joint Genome Institute. The work conducted by the US Department of Energy Joint Genome Institute, a US Department of Energy Office of Science User Facility, is supported by the Office of Science of the US Department of Energy under Contract DE-AC02-05CH11231. The US Department of Energy Joint BioEnergy Institute (www.jbei.org) is supported by the US Department of Energy, Office of Science, Office of Biological and Environmental Research, through Contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the US Department of Energy.

1. Samson RA, et al. (2014) Phylogeny, identification and nomenclature of the genus Aspergillus. *Stud Mycol* 78:141–173.
2. Nierman WC, et al. (2005) Genomic sequence of the pathogenic and allergenic filamentous fungus Aspergillus fumigatus. *Nature* 438:1151–1156.
3. Moore G, Mack B, Beltz S (2015) Draft genome sequences of two closely-related aflatoxigenic Aspergillus species obtained from the Ivory Coast. *Genome Biol Evol* 8:729–732.
4. Frisvad JC, Larsen TO (2016) Extrolites of Aspergillus fumigatus and other pathogenic species in Aspergillus section Fumigati. *Front Microbiol* 6:1485.
5. Peláez T, et al. (2013) Invasive aspergillosis caused by cryptic Aspergillus species: A report of two consecutive episodes in a patient with leukaemia. *J Med Microbiol* 62: 474–478.
6. Hoffmeister D, Keller NP (2007) Natural products of filamentous fungi: Enzymes, genes, and their regulation. *Nat Prod Rep* 24:393–416.
7. Inglis DO, et al. (2013) Comprehensive annotation of secondary metabolite biosynthetic genes and gene clusters of Aspergillus nidulans, A. fumigatus, A. niger and A. oryzae. *BMC Microbiol* 13:91.
8. Frisvad JC, Larsen TO (2015) Chemodiversity in the genus Aspergillus. *Appl Microbiol Biotechnol* 99:7859–7877.
9. Perrin RM, et al. (2007) Transcriptional regulation of chemical diversity in Aspergillus fumigatus by LaeA. *PLoS Pathog* 3:e50.
10. Palmer JM, Keller NP (2010) Secondary metabolism in fungi: Does chromosomal location matter? *Curr Opin Microbiol* 13:431–436.
11. Brakhage AA, Schroeckh V (2011) Fungal secondary metabolites: Strategies to activate silent gene clusters. *Fungal Genet Biol* 48:15–22.
12. Osbourn A (2010) Secondary metabolic gene clusters: Evolutionary toolkits for chemical innovation. *Trends Genet* 26:449–457.
13. Yu J, et al. (2004) Clustered pathway genes in aflatoxin biosynthesis. *Appl Environ Microbiol* 70:1253–1262.
14. Cary JW, Ehrlich KC, Beltz SB, Harris-Coward P, Klich MA (2009) Characterization of the Aspergillus ochraceoroseus aflatoxin/sterigmatocystin biosynthetic gene cluster. *Mycologia* 101:352–362.
15. Grigoriev IV, Martinez DA, Salamov AA (2006) Fungal genomic annotation. *Appl Mycol Biotechnol* 6:123–142.
16. Machida M, et al. (2005) Genome sequencing and analysis of Aspergillus oryzae. *Nature* 438:1157–1161.
17. Wortman JR, et al. (2006) Whole genome comparison of the A. fumigatus family. *Med Mycol* 44:S3–S7.
18. Mondo SJ, et al. (2017) Widespread adenine N6-methylation of active genes in fungi. *Nat Genet* 49:964–968.
19. Qi J, Luo H, Hao B (2004) CVTree: A phylogenetic tree reconstruction tool based on whole genomes. *Nucleic Acids Res* 32:W45–W47.
20. Zuo G, Li Q, Hao B (2014) On K-peptide length in composition vector phylogeny of prokaryotes. *Comput Biol Chem* 53:166–173.
21. Peterson SW (2008) Phylogenetic analysis of Aspergillus species using DNA sequences from four loci. *Mycologia* 100:205–226.
22. Kocsubé S, et al. (2016) Aspergillus is monophyletic: Evidence from multiple gene phylogenies and extrolites profiles. *Stud Mycol* 85:199–213.
23. Mitchell A, et al. (2015) The InterPro protein families database: The classification resource after 15 years. *Nucleic Acids Res* 43:D213–D221.
24. Hong S-B, Go S-J, Shin H-D, Frisvad JC, Samson RA (2005) Polyphasic taxonomy of Aspergillus fumigatus and related species. *Mycologia* 97:1316–1329.
25. Delcher AL, Phillippy A, Carlton J, Salzberg SL (2002) Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res* 30:2478–2483.
26. Kurtz S, et al. (2004) Versatile and open software for comparing large genomes. *Genome Biol* 5:R12.
27. Delcher AL, et al. (1999) Alignment of whole genomes. *Nucleic Acids Res* 27: 2369–2376.
28. Khaldi N, et al. (2010) SMURF: Genomic mapping of fungal secondary metabolite clusters. *Fungal Genet Biol* 47:736–741.
29. Medema MH, et al. (2015) Minimum information about a biosynthetic gene cluster. *Nat Chem Biol* 11:625–631.
30. Frisvad JC, Samson RA (1990) Chemotaxonomy and morphology of Aspergillus fumigatus and related species. *Modern Concepts in Penicillium Apergillus Classification*, pp 201–208.
31. Tamiya H, et al. (2015) Secondary metabolite profiles and antifungal drug susceptibility of Aspergillus fumigatus and closely related species, Aspergillus lentulus, Aspergillus udagawae, and Aspergillus viridinutans. *J Infect Chemother* 21:385–391.
32. Nielsen KF, Månsson M, Rank C, Frisvad JC, Larsen TO (2011) Dereplication of microbial natural products by LC-DAD-TOFMS. *J Nat Prod* 74:2338–2348.
33. Rank C, et al. (2008) Novofumigatonin, a new orthoester meroterpenoid from Aspergillus novofumigatus. *Org Lett* 10:401–404.
34. Guo C-J, et al. (2012) Molecular genetic characterization of a cluster in A. terreus for biosynthesis of the meroterpenoid terretonin. *Org Lett* 14:5684–5687.
35. Okuyama E, Yamazaki M, Katsube Y (1984) Fumigatonin, a new meroterpenoid from Aspergillus fumigatus. *Tetrahedron Lett* 25:3233–3234.

36. Itoh T, et al. (2010) Reconstitution of a fungal meroterpenoid biosynthesis reveals the involvement of a novel family of terpene cyclases. *Nat Chem* 2:858–864.
37. Latgé JP (1999) Aspergillus fumigatus and aspergillosis. *Clin Microbiol Rev* 12:310–350.
38. Aalberse RC, Akkerdaas J, van Ree R (2001) Cross-reactivity of IgE antibodies to allergens. *Allergy* 56:478–490.
39. Valiante V, Macheleidt J, Föge M, Brakhage AA (2015) The Aspergillus fumigatus cell wall integrity signaling pathway: Drug target, compensatory pathways, and virulence. *Front Microbiol* 6:325.
40. Rementeria A, et al. (2005) Genes and molecules involved in Aspergillus fumigatus virulence. *Rev Iberoam Micol* 22:1–23.
41. Hong S-B, et al. (2010) Re-identification of Aspergillus fumigatus sensu lato based on a new concept of species delimitation. *J Microbiol* 48:607–615.
42. Ehrlich KC, Yu J, Cotty PJ (2005) Aflatoxin biosynthesis gene clusters and flanking regions. *J Appl Microbiol* 99:518–527.
43. de Guzman FS, et al. (1994) Ochrindoles A-D: New bis-indolyl benzenoids from the sclerotia of Aspergillus ochraceus NRRL 3519. *J Nat Prod* 57:634–639.
44. Frisvad JC, Frank JM, Houbraken JAMP, Kuijpers AFA, Samson RA (2004) New ochratoxin A producing species of Aspergillus section Circumdati. *Stud Mycol* 50:23–43.
45. Balibar CJ, Howard-Jones AR, Walsh CT (2007) Terrequinone A biosynthesis through L-tryptophan oxidation, dimerization and bisprenylation. *Nat Chem Biol* 3:584–592.
46. Richards M, Bird AE, Munden JE (1969) Chlorflavonin, a new antifungal antibiotic. *J Antibiot* 22:388–389.
47. Burns MK, Coffin JM, Kurobane I, Vining LC (1979) Biosynthesis of chlorflavonin in Aspergillus candidus : A novel fungal route to flavonoids. *J Chem Soc Chem Commun* 426–427.
48. Hashimoto M, Nonaka T, Fujii I (2014) Fungal type III polyketide synthases. *Nat Prod Rep* 31:1306–1317.
49. Austin MB, Noel JP (2003) The chalcone synthase superfamily of type III polyketide synthases. *Nat Prod Rep* 20:79–110.
50. Varga J, Frisvad JC, Samson RA (2007) Polyphasic taxonomy of Aspergillus section Candidi based on molecular, morphological and physiological data. *Stud Mycol* 59: 75–88.
51. Vaillancourt FH, Yeh E, Vosburg DA, O'Connor SE, Walsh CT (2005) Cryptic chlorination by a non-haem iron enzyme during cyclopropyl amino acid biosyn. *Nature* 436: 1191–1194.
52. Kirner S, et al. (1998) Functions encoded by pyrrolnitrin biosynthetic genes from Pseudomonas fluorescens. *J Bacteriol* 180:1939–1943.
53. Xu X, et al. (2014) Identification of the first diphenyl ether gene cluster for pestheic acid biosynthesis in plant endophyte Pestalotiopsis fici. *ChemBioChem* 15:284–292.
54. Fullone MR, et al. (2012) Insight into the structure-function relationship of the non-heme iron halogenases involved in the biosynthesis of 4-chlorothreonine–Thr3 from Streptomyces sp. OH-5093 and SyrB2 from Pseudomonas syringae pv. syringae B301DR. *FEBS J* 279:4269–4282.

3.2 Manuscript II – Comparative genomics of section *Flavi*

The *Aspergillus* section *Flavi* belongs to subgenus *Circumdati*. The section contains species extremely important to our society including the industrial workhorse *A. oryzae*, the food fermenting *A. sojae* and the crop spoilers and toxin producers *A. flavus* and *A. parasiticus*. Besides these (in)famous species, the section also includes at least 19 other less studied species.

We have made a genomic comparisons of 23 members of the section, where we *de novo* sequenced 19 species. The species already sequenced were all from the same clade end hence only covered a small part of the diversity of the section. With our additional species we have a wide coverage of the section and even more species from the most studied clade giving an even higher resolution. Our analysis has shown that section *Flavi* have certain general characteristics, including large genomes which is reflected in the high number of predicted secondary metabolite gene clusters and carbohydrate active enzymes (CAZymes). These features make the *Flavi* section highly interesting for future genome mining.

With whole genome sequences covering the entire section, it has been possible to investigate issues such as phylogeny and toxin production to gain new insights. The phylogeny was for instance investigated based on 200 genes revealing evolutionary patterns that have not been seen before concerning the relatedness of species within the extensively studied *A. flavus* clade. The distribution of secondary metabolite gene clusters including known toxins such as aflatoxin has been investigated to show the potential across the section and compared with experimental data.

The findings presented in Manuscript II illustrates even more uses of comparative genomics both in terms of what analysis that can be used and what insights we can gain.

Manuscript II will be submitted to Genome Biology. The additional files can be found in Appendix B.

RESEARCH

Friends and foes – A comparative genomics study of 23 Aspergillus species from section Flavi

Inge Kjærbølling[1], Tammi Vesth[1], Jens C. Frisvad[1], Jane L. Nybo[1], Sebastian Theobald[1], Sara Kildgaard[1], Thomas I. Petersen[1], Alan Kuo[2], Atsushi Sato[3], Ellen K. Lyhne[1], Martin E. Kogle[1], Ad Wiebenga[6], Roland S. Kun[6], Ronnie Lubbers[6], Miia R. Mäkelä[7], Alicia Clum[2], Anna Lipzen[2], Asaf Salamov[2], Chew Y. Ngan[2], Chris Daum[2], Jennifer Chiniquy[2], Kerrie Barry[2], Kurt LaButti[2], Sajeet Haridas[2], Blake A. Simmons[4], Jon K. Magnuson[4], Bernard Henrissat[8], Thomas O. Larsen[1], Ronald P. de Vries[6], Igor V. Grigoriev[2,10], Uffe H. Mortensen[1], Masayuki Machida[5], Scott E. Baker[4,9] and Mikael R. Andersen[1*]

Abstract

Background: Section *Flavi* encompasses several extremely important species both harmful and beneficial. The best known are most likely *A. oryzae* used in food fermentation and enzyme production and *A. flavus*, a food and feed spoiler and mycotoxin producer. As part of the whole-genus *Aspergillus* sequencing project we have generated 19 genomes *de novo* spanning the breadth of the *Flavi* section. We have made comparisons of 31 fungal genomes including 23 species from the *Flavi* section in order to investigate the genomic diversity of this important section.

Results: Based on the genomes, we have predicted 13,759 CAZymes and 1,972 secondary metabolite gene clusters and shown that the *Flavi* species in general have a high number of both CAZymes and secondary metabolite gene clusters compared to other *Aspergillus* species. The CAZy content was compared with growth experiments on 35 media showing that the genetic variations are not necessarily reflected in the growth patterns. The secondary metabolite gene clusters were divided into families and coupled to compounds, allowing predictions of 20 compound families in 31 species. Focusing on the highly toxic aflatoxin gene cluster, we have shown that it is conserved in 14 section *Flavi* species and that a very short version is found in *A. caelatus*, and most likely a complete loss has happened in *A. tamarii*. By coupling the cluster predictions to chemical data of the *Flavi* species we have been able to identify a putative miyakamide biosynthetic gene cluster demonstrating one of the versatile possibilities unlocked from this data.

Conclusion: In summary we have generated a highly valuable and versatile resource and used it to demonstrate the diversity and similarities of the *Flavi* section both in terms of CAZy content and secondary metabolite potential.

Keywords: Aspergillus; Comparative Genomics; Secondary Metabolism

Background

Aspergillus section *Flavi* encompasses a large number of species, many of whom have a significant impact on human life: Some species (e.g. *A.oryzae* and *A. sojae*) are extensively used in food fermentation for the production of sake, miso, soy sauce, and other fermented foods. Moreover, *A. oryzae* is also used industrially for enzyme and secondary metabolite production [36, 46, 65]. In contrast, other *Flavi* species (e.g. *A. flavus* and *A. paraciticus*) are notorious for producing highly toxic fungal compounds, the aflatoxins, in addition to infecting and damaging crops [25, 34, 83]. Furthermore, *A. flavus* has the capability of infecting immunocompromised humans, and is now the second leading species associated with human aspergillosis [27, 38]

Besides these well-studied species, the section also includes several less known species that like their (in)famous relatives includes both beneficial and harmful properties. There are potential positive impacts from the producers of bioactive compounds (antiinsectant N-alkoxypyridone metabolite, leporin A, from *A. leporis*; an antibiotic with antifungal activity, avenaci-

*Correspondence: mr@bio.dtu.dk
[1]Department of Biotechnology and Bioengineering, Technical University of Denmark, Søltoft Plads 223, 2800 Kongens Lyngby, Denmark
Full list of author information is available at the end of the article

olide, from *A.avenaceus*) and enzymes producers (including amylases, proteases and xylanolytic enzymes in *A.tamarii* and pectin degrading enzymes in *A. alliaceus*). On the harmful side, plant pathogens (*A. alliaceus* on onion bulb, *A. nomius* on nuts, seeds and grains) and toxin producers (ochratoxin from *A. alliaceus*, aflatoxin from *A. nomius*) are also found among these less studied *Flavi* species [79].

Given the high impact that species of Section *Flavi* have in industrial enzyme production, food fermention, toxin production and crop disease, it is important to examine the full potential of the section to assess the potential alternative species for industrial use, novel bioactive compound production and as a source of useful enzymes while also characterizing each genome for the purpose of mitigating food spoilage, mycotoxin production and pathogenicity.

At the onset of this project, whole geneome sequence was available for five species from section *Flavi* (*A. oryzae, A. flavus, A. sojae, A. bombycis, and A. paraciticus*) [43, 45, 56, 62, 65]. They all belong to the *A.flavus* clade which is closely related and only covers a small part of the total section. In this study, as part of the *Aspergillus* Genus Sequencing Project, we have generated genome sequence for 18 additional species plus an additional *A. parasiticus* isolate, permitting genomic comparisons across 23 members of Section *Flavi*.

Comparing genomes across a section-wide representation allows for examination of important features for food fermentation and exploration of enzyme and bioactive compound potential. We gain insights into the core attributes of the section as well as the differences and diversity across the section, which will give a better understanding of the widely used species and an appreciation of the potential of the section with many unexplored resources. The understanding gained from this study are useful for 1) exploring novel enzymes and secondary metabolites, 2) optimizing food fermentation and industrial use and 3) improve food and feed protection and control. We have applied advanced comparative genomics to identify core, species-specific and section and clade specific genes. From our analysis we developed a deeper understanding of the diversity and similarities within the section and of each species.

Results and Discussion

Analysis of 19 newly sequenced genomes show large genome size in section *Flavi*

In this study we present the whole genome sequences of 19 species from *Aspergillus* section *Flavi* (Figure1 panel B). The analysis of these genomes also include previously sequenced section *Flavi* species (*A. oryzae, A.flavus, A. sojae, A. bombycis* [45, 56, 62, 65]). During the time of this project two additional *Flavi* genomes have been published, *A. nomius* and *A. arachidicola* [55, 57] which have not been used in this analysis instead we have used our own sequenced strains of these species. In order to assess the quality of the data and generate a picture of the genome diversity of the *Flavi* section we investigated the quality of the genome assemblies and compared the genome attributes such as genome size, GC content and number of predicted proteins. The results are displayed in Figure 1C.

The genome assemblies generated for this study are of similar quality with 13 out of the 18 genomes assembled into fewer than 500 scaffolds, Figure 1C (column 5). *A. coremiformiis* has the highest number of scaffolds, 2,728. With such a high number of scaffolds we were concerned with the quality of the genome. However the percentage of core genes from BUSCO (Benchmarking Universal Single-Copy Orthologs) [69] is 99.78% lacking only a phosphomannomutase, suggesting that the gene content is of appropriately high quality despite the large number of scaffolds.

The genome sizes of *Aspergillus* section *Flavi* are large with a mean of 37.96 Mbp compared to 31.7 Mbp for other representative Aspergilli, see Figure 1C (column 1). Notably, in section *Flavi* there is one particular exception — *A. coremiformiis* — which is only 30.1 Mbp, also reflected in the number of predicted proteins ranging from 9078 to 14216 for *A. coremiformiis* and *A. novoparasiticus* respectively. The large genome sizes compared with other *Aspergilli* have been reported for *A.oryzae* [19] and the data presented here suggests this is a trait shared by most species in the *Flavi* section. Functional annotation covers 68-81% of the predicted proteins in section *Flavi*.

From our investigations we confirm that the genomes are of high quality and that a section *Flavi* specific trait is large genomes both in terms of size (Mbp) and number of predicted proteins (with *A. coremiformiis* as one major exception).

Phylogeny of section *Flavi*

Next we examined the evolutionary relationships and relatedness of section *Flavi* based on genome sequences and predicted proteins. To do this we constructed a maximum likelihood phylogenetic tree based on 200 monocore genes (genes having exactly one homolog in each genome), Figure 1A. The tree shows that section *Flavi* is a monophyletic group. The clades within section *Flavi* correspond to what has been seen for a phylogenetic tree based on the β-tubulin gene with the *A. tamarii* clade as neighbour to the *A. flavus*, followed by the *A. nomius, A. togoensis, A. alliaceus, A.*

leporis and *A. avenaceus* clade [79]. *A. bertholletius* was not included in the study by Varga et al., but Taniwaki et al. have shown that it falls in between the *A. nomius* and *A. togoensis* clades in its own distinct clade which is confirmed here [75]. The species from section *Circumdati* and *Terrei* are closest related to the *Flavi* section followed by the *Nigri, Nidulantes*, and *Fumigati* section which was also seen by Kocsube et al. [37].

The support of the branching within the tree is generally high, with most branch points supported by 100 out of 100 bootstraps. This also includes the distinction between section *Flavi* and the neighbouring sections. Within section *Flavi* the only branch below 80 bootstraps is between *A. flavus* and *A. sojae*. It is indeed also surprising that *A. sojae* is found closest to *A. flavus* since *A. sojae* is perceived as a domesticated version of *A. parasiticus*, while *A. oryzae*, perceived as a domesticated version of *A. flavus* is not next to it in the tree [39, 79, 84]. There have been reports suggesting that *A. oryzae* descended from an ancestor that was the ancestor of *A. minisclerotigenes* or *A. parvisclerotigenus* based on the region neighbouring the cyclopiazonic acid biosynthetic gene cluster [8] which may be supported by this tree where *A. oryzae* is between *A. minisclerotigenes* and *A. parvisclerotigenus*.

It is worth noting that within the *A. flavus* clade, the species are very closely related it can therefore be difficult to resolve the inter-species relationships. In an attempt to increase the resolution and resolve the *A. flavus* clade we created phylogenetic trees based on 300 and 500 monocore genes, see additional file Figure B1 and B2. In each of the generated trees only *A. sojae* changes position while the other *A. flavus* clade species have steady positions. The bootstrap values for *A. sojae* are 37 and 32 out of 100 for the tree based on 300 and 500 sequences respectively. In none of the trees *A. sojae* is closest to *A. parasiticus*, but rather in between the group with *A. flavus* and *A. parasiticus*. One explanation for the unstable behavior of *A. sojae* could be that it is a hybrid species, or that the process of domestication has had significant effects and a different evolutionary pressure. To investigate the speculation that *A. sojae* could be a hybrid species, which would explain why it is so difficult to place in the phylogenetic tree, we created single gene trees for 18 of the monocore protein families (Additional file B3). We would expect that some genes are closely related to one parent while other genes are closely related to the other parent in a hybrid species. For 6 of the single gene trees we see that *A. sojae* is closest related to *A. flavus*, for 6 other we see *A. sojae* is closer to *A. parasiticus*, in 2 trees *A. sojae* is in between and the final 4 trees are inconclusive due to lack of resolution. This further supports the hypothesis that *A. sojae* potentially could be a hybrid species.

In summary, we show that section *Flavi* is a monophyletic group and that the previously defined clades are reflected by the groupings in phylogenies built on 200–500 genes. However it is not possible to resolve the phylogeny within the *A. flavus* clade based on whole genome phylogeny based on 200–500 monocore genes.

Examination of core, section-specific and species-specific genes

In the view of the genome attributes from the previous section we wanted to investigate the common traits of section *Flavi* and the diversity as individual species-specific characteristics in order to generate an understanding of core feature shared by all section *Flavi* species and to identify variations and specific features important for their various uses and purposes such as food fermentation, plant and human pathogenicity.

To do this we constructed families of homologous proteins across the species found in Figure 1 using an existing pipeline [80]. This approach makes it possible to identify homologs within and across the species, we use the term homolgs since we are not doing functional analysis of these genes and therefore we cannot confidentially disticguish between orthologs and paralogs. Using these homologous protein families we identified 1) the core genome - protein families with at least one member in all species. This is expected to be essential functions such as those important for growth and primary metabolism. 2) Clade and section specific genes - genes that have homologs in all members of the group but not with any other species. These genes are expected to be involved in phenotypical traits specific for that group, it could be secondary metabolite genes or specific carbon utilization. 3) Species-specific genes - genes without homologs in any other species in the comparison. Species-unique genes are presumed to be possible gene annotation errors or encompassing phenotypical traits unique to one species important for speciation and adaptation [80].

With regard to core genes, it is important to note that depending on which species are included in the construction of the families, the core will change: the closer related the species are, the bigger the core will be as those species will have more functions in common. Thus the core of the set including all 31 species in this dataset is 2,082 protein families, whereas the core including only the 29 *Aspergillus* species is 3,853 and for section *Flavi* alone, it is 4,903 protein families. The *Aspergillus* core ranges from 38.04% for *A. novoparasiticus* to 50.83% for *A. coremiiformis* of the total genome while the *Flavi* core ranges from 47.35-63.43%

for *A. novoparasiticus* and *A. coremiiformis* respectively. Hence, about half of the section *Flavi* genomes are varied when comparing across the section.

By coupling the protein families to the phylogenetic tree, it is possible to identify section and clade specific genes and species unique genes. There are 92 section *Flavi*-specific protein families. The number of clade-specific families ranges from 27 (*A. flavus* clade), 28 (A. tamarii clade) to 54 for the *A. nomius* clade which is low compared with section *Nigri* examined previously [80]. The fact that there are few clade-specific genes could indicate that some clades are closely related and that species from two different clades in section *Flavi* might share many features. The *A. flavus* and *A. tamrii clade* for instance shares 22 protein families which is nearly as many as the clade specific families they have.

Lastly, the species-specific genes for section *Flavi* species ranges from 2,181 to 166 in *A. leporis* and *A. sojae* respectively. The number of species-specific genes in *A. sojae* is very low, compared to an earlier study of section *Nigri* including 6 different strains of *A. niger* where the number of strain-specific genes ranged from 182 for NRRL3 to 1015 for CBS 513.88 but most had around 3-400 strain-specific genes [80]. While this raises the question whether *A. sojae* is or is not its own species, it should be noted that the *A. sojae* genome was annotated with different gene modeling pipeline than the genomes we generated for this study and could therefore influence these numbers. The species question has been discussed previously concerning the relationship between *A. flavus* and *A. oryzae* and it seems that the argument of practicality and food safety is a factor when regarding them as separate species for regulatory reasons rather than evolutionary or phylogentic reasons [20, 24, 35].

As mentioned above, the species-unique genes are expected to retain a part of what sets a species apart from other species [80] making it interesting to investigate these genes in more detail. We examined the functions to investigate if they live up to the expectation and the statistical significance to see if there is an enrichment of specific functional domains in species unique genes.

We examined the functions of the species unique genes using InterPro, GO and KOG annotations [18, 22, 23, 76]. The portion of species-specific genes with a functional annotation was 20, 12 and 9 percent for InterPro, GO and KOG respectively. In total, 21% of the species-specific genes had an annotation. We will focus on InterPro annotations since it covers the most (most common GO and KOG annotations, Additional file, Figure B5 and B6). For InterPro, the most common functions range from transcription factor, protein kinase, transporter to P450 (Figure Additional file, Figure B4). These functions fit well with our expectation of proteins that could be involved in phenotypic traits.

For the most common IPR, GO and KOG domains, we investigated whether these functions are more often associated with species-specific genes. We used Fisher's exact test to investigate if there is a significant association between the annotated functions and species-specific genes. The majority of the most common domains have p-values below 0.005 or 0.00001 which are indicated by stars in the figures of Additional file Figure B4, B5 and B6. Looking at all the functions found in species unique genes approximately 43% of the InterPro domains are enriched in species unique genes with p-values below 0.01, indicating that there is an enrichment of certain functional domains in species unique genes.

Species-specific and secondary metabolite genes are localized towards the the subtelomeric regions

In a study by Galagan et al. it was shown that the subtelomeric sequences are associated with extensively rearranged regions lacking synteny when comparing *A. nidulans, A. oryzae* and *A. fumigatus* [19]. This phenomenon has also been described for mammals, nematodes and yeasts [15]. A comparative genomic hybridization (CGH) study of *A. fumigatus* Af293 with two other *A. fumigatus* strains and three closely related species showed that there was a bias towards sub-telomeric regions for the unique, diverged or missing genes [61]. Another feature enriched in the subtelomeric regions is the presence of predicted secondary metabolite gene clusters (SMGCs) seen for *A. nidulans* and *A. fumigatus*. It was thus suggested that the sub-telomeric region promotes species specific evolution of the secondary metabolite genes (SMGs) and that the SMGs play a role in the ecology [19].

We therefore examined the location of species-unique genes and predicted secondary metabolite clusters in order to assess the potential over-representation of these genes in telomeric regions. We used the nearly completely assembled genome (distributed on 8 chromosomes plus 3 scaffolds) of *A. oryzae* as the reference for analysis of gene location, Figure 3. We calculated the density of the genes across the genome (Figure 3, black graph), identified the core genes and mapped them to the genome (grey dots), shown below the chromosomes in Figure 3. The 696 *A. oryzae* species-specific genes and the 456 predicted SMG were identified and mapped to the genome shown above the chromosomes, Figure 3. Based on the density of the total number of genes, the percentage of the unique and the SMG were calculated and plotted above the chromosome as graphs.

From our plot it is clear that the core genes are found spread across all the chromosomes, but with a lower concentration at the sub-telomeric ends. In the sub-telomeric ends, unique and SMG are definitely found but some are also spread across the genome. The species specific and clade specific genes from *A. oryzae* were also mapped to the genome but no clear pattern was observable. To examine whether or not the pattern of the unique and SMGs is statistically significant we performed Fisher's exact test which showed that both unique (p-value = 7.266e-07) and SMGs (p-value < 2.2e-16) are enriched towards the sub-telomeric regions (100 kbp from the chromosomal ends) supporting previous work by Galagan et al. and Nierman et al. [19, 61]. The fact that the species-specific genes are not randomly distributed argues against the possibility that species-specific genes are simply annotation or gene modeling errors, therefore indicating that they are, indeed, legitimate genes. The distribution of the species-specific genes could indicate the existence of a mechanism where new genes enter the genome at telomeric regions with higher frequency, or it could be an indication of a counter selection preventing new genes in conserved regions potentially because those changes kills the host in most cases or it could be a combination of both.

Analysis of syntenic and non-syntenic regions in section *Flavi*

Another factor to consider when analysing genome location is syntenic regions and non syntenic regions. It has been shown that the *A. oryzae* genome has a mosaic pattern of syntenic and non syntenic regions with more non-syntenic regions proximal to telomeric ends [36, 46]. We examined the synteny across the *Flavi* section using *A. oryzae* as reference and compared with the section *Flavi* genomes plus *A. nidulans* and *A. fumigatus*. The number of syntenic genes and the percentage of the *A. oryzae* genome can be seen in Table 1. *A. oryzae* shares the highest synteny with *A. minisclerotigenes* with 75.59% of the genes being syntenic followed by 72.20% to *A. flavus*. This is another indication that *A. oryzae* is closest related to *A. minisclerotigenes* and not *A. flavus*.

Shared syntenic genes are illustrated on Figure 4, where the genes syntenic in all species in a group (*A. flavus* clade; *A. flavus* and *A. tamarii*; *A. flavus, A. tamarii* and *A. nomius*; *Flavi* section; *Flavi* section, *A. nidulans* and *A. fumigatus*) are shown. The species within the *A flavus* clade shares a high degree of synteny which is conserved across most of the genome. When more distantly related species are included the syntenic regions are reduced.

In general, there are fewer regions of synteny towards the telomeric ends corresponding with earlier studies made with three species from different clades (*A. nidulans*, *A. fumigatus* and *A. oryzae*) [36, 46] . The more distant the compared species, the more pronounced the trend. We also observe that some chromosomes (especially chromosome 1 and 2) has a very high degree of conserved synteny across all the species compared, while others (such as chromosome 6 and 8) have a much lower syntenic conservation.

A. oryzae genes that are non-syntenic in all the species in this study are marked below the chromosome in black, Figure 4. They are as expected based on the analysis above and previous studies found in a higher density towards the telomeric ends. More surprisingly there are some blocks of non-syntenic genes found with very high density at some specific regions, one on chromosome 4, 6 and 8. One explanation for these non syntenic blocks could be horizontal gene transfer (HGT), another explanation could be gene shuffling or de novo gene formation of *A. oryzae* specific genes potentially by a major event of gene duplication followed by divergence.

HGT was investigated using BLASTp to examine the best hits in the NCBI non-redudant database. Recent HGT are expected to have high identity with another group of species where it would have been transferred from, and not be found in the closely related species. None of the genes within these blocks showed sign of recent HGT (data not shown). The non-syntenic block genes were compared with the identified *A. oryzae* unique genes. Only 23 of the 80 genes in the non-syntenic blocks were *A. oryzae* unique genes. It thus seems likely that these non-syntenic blocks are caused by significant rearrangements and the emergence of *A. oryzae* unique genes.

Taken together, these observations — some very conserved chromosome and some highly rearranged non syntenic blocks — could indicate an evolutionary pressure for stability in some regions while other regions are frequently subject to gene shuffling and rearrangements, i.e. rearrangement hot spots.

Section *Flavi* is a rich source of Carbohydrate Active Enzymes

Carbohydrate-Active enZymes (CAZymes) are essential for what carbon sources a species can degrade and utilize. The *Aspergillus* genus is known to have a high number of these enzymes and a wide span and hence the ability to degrade and convert a broad range of plant biomass [13]. Aspergilli are considered generalists concerning carbon utilization since they are capable of degrading various plant biomass and are thus not specialized on one thing [4, 12]. Within section *Flavi* this is mainly described for *A. oryzae* [3, 36, 46] and to a lesser extent for *A. flavus* [4, 31, 47, 59, 70]

and A. sojae [29, 51], while only incidental studies have been performed with other species of this group [11, 28, 49, 58, 64, 66, 68], often describing production or characterization of certain CAZyme activity or protein, respectively.

To investigate the genomic diversity and enzyme potential of section *Flavi*, we used the Carbohydrate-Active Enzymes database (CAZy) to predict the CAZyme content in each of the genomes. A total of 13,759 CAZymes were predicted within the 23 *Flavi* species with an average of 598 per species which is higher than the other Aspergilli in the set having an average of 508 per species, Figure 5A. *A. parasiticus* has the highest CAZyme content with 663 while *A. coremiformiis* has the lowest with 383.

Each of the identified CAZymes belongs to one of the seven classes of enzyme activity. The most common CAZyme class is composed of the Glycoside hydrolases (GH) with more than 250 proteins per species followed by Glycosyltransferases, Auxiliary activities, Carbohydrate-binding molecules, Carbohydrate esterases, Polysaccharide lysases and Distant plant expansins. Comparing the number of proteins within each CAZyme class for the clades within section *Flavi*, glycoside hydrolases has a higher number of proteins in clade *A. flavus* followed by *A. tamarii* and *A. nomius*, Figure 5B. The similar pattern is found for Auxiliary activities where the highest number is found in *A. flavus* followed by *A. nomius* and *A. tamarii*. For Glycosyltransferases the highest number of proteins is found in clade *A. nomius* with slightly lower but similar number of proteins in the rest of section *Flavi*. For carbohydrate esterases the number of proteins is very similar across the groupings.

Comparing the diversity, the number of subgroups across the CAZyme classes it is quite similar across clades for most of the classes except Glycoside hydrolases where the *A. flavus, A. tamarii* and *A. nomius* clades had higher diversity compared to the rest, Figure 5B. Polysaccharide lysases have a slightly higher diversity for the *A. flavus* and *A. tamarii* clade.

The different clades of section *Flavi* contain variable CAZyme gene content, but this does not seem to specifically effect their ability to degrade plant biomass

To evaluate the actual carbon utilization ability across section *Flavi* and compare it to other fungal species, we performed growth profiling of 31 species (29 Aspergilli, including 23 species from section *Flavi*) on 35 plant biomass-related substrates, Figure 6 (Additional file Figure B7 and Table B1) and compared this to the CAZy gene content prediction related to plant biomass degradation (Additional file Table B2). D-Glucose resulted in the best growth of all monosaccharides for all species and was therefore used as an internal reference for growth, Additional Figure B7. Growth on other carbon sources was compared to growth on D-glucose and this relative difference was compared between the species. While in a previous study the variation in growth between distantly related Aspergilli could be linked to differences in CAZy gene content in their genomes [13], this was not the case in a recent study addressing species of the section *Nigri* of *Aspergillus* [80]. The CAZyme sets of section *Flavi* related to plant biomass degradation are overall highly similar (Figure 6), especially when comparing species of the same clades, but there are some noteworthy exceptions. *A. coremiiformis* has a strongly reduced gene set compared to all other members of section *Flavi*, including its closest relatives (*A. leporis* and species of the *A. alliaceus* clade), as well as to the other Aspergilli, and is similar in CAZyme content to *Penicillium digitatum*.

The *A. alliaceus* and *A. nomius clades* together with *A. bertholletius* are highly similar in CAZyme content, but contain a somewhat lower number than the *A. tamarii* and *A. flavus* clades (Figure 6). The *A. tamarii* and *A. flavus* clades are also highly similar, with the exception of *A. coelatus* (*A. tamarii* clade) that has even lower number of CAZymes than the *A. nomius* clade. When looking more closely at number of genes in individual CAZy families in section *Flavi*, Additional file Table B2, some differences can be observed that may affect their hydrolytic abilities. *A. alliaceus* is expanded in specific xylanolytic (GH11 endoxylanase) and xyloglucanolytic (GH12 xyloglucan-active endoglucanase) genes, while expansion of the number of genes in the cellulolytic family GH45 in this species is shared with another members of the *A. flavus* clade (*A. parasiticus*) as well as *A. coremiiformis* (clade *A. togoensis*). Differences are also observed between the clades of section *Flavi* (Additional file Table B2). Clade *A. togoensis* has a reduced set of xylanolytic and xyloglucanolytic genes, but this is not reflected in the growth of these species on xylan. Two xyloglucanolytic families (GH29 α-fucosidase and GH74 xyloglucan-active endoglucanase) were absent in all species of section *Flavi*, but present in other Aspergilli. In contrast, expansion of several xylanolytic families was observed in specific clades. GH115 (α-glucuronidase) is expanded in clades *A. flavus, A. tamarii* and *A. nomius*. Accordingly, xylanolytic enzymes or activity have been reported from several species from these clades [17, 28, 32, 49, 54, 64, 68, 71, 72]. GH62 (arabinoxylan arabinofuranohydrolase) was expanded in clade *A. leporis*. Clades *A. leporis* and *A. avenaceus* were the only clades from section *Flavi* that had genes of CE15 (glucuronoyl esterase), which were also found in *Aspergillus* species outside section *Flavi*.

The galactomannan degrading ability was nearly fully conserved in section *Flavi*, but interestingly growth on guar gum that consists mainly of galactomannan was variable between the species. We recently demonstrated that guar gum contains small amounts of other sugars that can trigger the production of enzymes with preferred activity towards other polysaccharides than galactomannan (R.P. de Vries, personal communication, manuscript in preparation), suggesting that the variation on growth is more likely due to regulatory differences in gene expression. Similarly, the reduced amylolytic ability of clades *A. togoensis* and *A. avanaceus* did not result in reduced growth on starch or maltose.

Large variation of the presence of inulinolytic genes (GH32) was observed among the species. Endoinulinase encoding genes were only present in clade *A. alliciaceus*, and two species of clade *A. flavus*. These species also contained an exoinulinase that was otherwise only found in three species from clade *A. tamarii*. The presence on endoinulinase resulted in better growth on inulin for most of those species, but surprisingly good growth was also observed for *A. coremiiformis* and *A. leporis*, which only contain invertases.

Variation was also observed in the number of pectinolytic genes. The most pronounced differences were the absence of PL11 (rhamnogalacturonan lyase) genes from most species of section Flavi, except for *A. avenaceus* and the expansion of GH78 (α-rhamnosidase) in clades *A. flavus* and *A. tamarii*. However, these differences and the smaller ones in other families did not result in large variation in growth on pectin. Interestingly though, some species, e.g. *A. oryzae, A. parasiticus, A. pseudonomius* and *A. sojae*, showed reduced growth on apple pectin compared to citrus pectin, but whether this is due to subtle changes in the pectinolytic ability or sensitivity to possible impurities in apple pectin is unclear at this point. Similarly, growth on different crude substrates was similar for most species, but variable for *A. oryzae, A. pseudonomius* and *A. sojae*. This variation may also be due to sensitivity to impurities in the substrates or to fine-tuning of gene expression resulting variably efficient enzyme mixtures.

As mentioned above, *A. coremiiformis* had a strong reduction in CAZyme content. While this was due to reduced numbers in some families, this species also specifically lacked a number of families related to pectin (GH54, GH88, PL1-PLY, PL4, PL9), xylan (GH62, GH67) and xyloglucan (GH31-AXL, GH95) degradation. Interestingly, this species showed better relative growth on xylan than most other species, while growth on other polysaccharides was mainly similar to that of the other species of section *Flavi*. This indicates that the reduced gene set of *A. coremiiformis* has not reduced its ability to degrade plant biomass. This could suggest that plant biomass degradation approach of *A. coremiiformis* is more similar to that found for *T. reesei* rather than *A. niger*, as *T. reesei* also has a reduced CAZy gene set, but produces the corresponding enzymes at very high level [50].

Growth on monosaccharides was largely similar between the species of section *Flavi* Figure 6 (additional file Figure B7 and Table B1). Growth of all species was equally good on D-glucose, D-fructose, D-mannose and D-xylose. This was also the case for L-ribose, except for *A. parvisclerotigenus, A. pseudotamatii, A. sergii* and *A. sojae*. Growth on L-arabinose was a slightly slower, and even more reduced on L-rhamnose and D-galacturonic acid, but also relatively similar across the species. Growth was poor on D-galactose, which is likely due to similar problems with uptake of D-galactose during germination as has been demonstrated for *A. niger* [16].

More obvious differences were present during growth on cellobiose and lactose. Except for *A. albertensis, A. alliaceus* and *A. pseudonomius*, most species grew poorly on cellobiose despite similar numbers of β-glucosidase encoding genes in most species Additional file Table B2. Similarly, only *A. arachidicola*, and to a lesser extent *A. albertensis* grew well on lactose, while the number of β-galactosidases in these species is similar to that of the other species.

Genomically the CAZy potential is largely conserved with some variations in copy numbers, but comparing with for instance the diversity found within secondary metabolism, the variation in CAZy potential is much smaller. The genomic potential and variations are not necessarily reflected in the growth. One explanation could be that that additional copies have evolved new functions or that the conditions tested does not trigger the expression of the enzymes. In conclusion, the CAZyme gene sets of section *Flavi* is largely conserved, with the exception of *A. cremiiformis*, but variation of growth in plant biomass related oligo- and polysaccharides was observed between them. It is therefore likely that as suggested previously [13], these differences are largely at the regulatory level.

Investigation of selected CAZy families important for food fermentation

We were particularly interested in enzymes belonging to the GH28 and GH13 families, as these are important for the food fermentation process and the quality of the product [77]. We were therefore interested in investigating these families and comparing the number and the relatedness of members of specific CAZyme subfamilies across the *Flavi* section. A phylogenetic tree was created of all members of GH28 from section

Flavi, Additional file Figure B8. The tree consists of 429 proteins, on average 18.7 per species. The most is found in *A. sergii* and *A. sojae* with 25 and 24 members respectively.

Within the tree there are different groupings. Five have members from all the 23 species (marked with blue squares). There are nine groupings missing one to four species (usually *A. coremiformiis* and *A. caelatus*), two groups have members of all *A. flavus, A. tamarii* and *A. nomius* clades. Lastly there are eight groupings containing two to 13 species which does not follow the phylogeny. It seems that the groupings containing all species are distributed throughout the tree often surrounded by groups missing some a few species and groupings with some species not following the phylogeny. In general species from clade *A. flavus* has a high number of GH28 members. *A. sojae* is known to have a high number of GH28 which is also seen here, but *A. sergii* has even higher. It could be interesting to investigate and see if this could be exploited either using *A. sergii* as a new species in food fermentation or as a source of novel enzymes.

Secondary Metabolism

Secondary metabolites (SMs) are important source of beneficial bioactive compounds and harmful toxins. In additon, SMs are thought to be involved in speciation [80]. The *Aspergillus* genus is known to produce a large number of SMs and the number of predicted secondary metabolite gene clusters (SMGCs) is even higher. The majority of predicted SMGCs are uncharacterized and therefore have the potential to produce a diversity of novel, bioactive compounds.

We examined the diversity and potential for SM production in section *Flavi*, both quantitatively in terms of numbers of clusters, and qualitatively in terms of the compounds these clusters could potentially produce. We utilized a SMURF-like SMGC prediction, followed by the creation of SMGC families (expected to produce the same compound or a derivative) and cluster dereplication [78].

High diversity of section Flavi secondary metabolism

To quantitatively assess the potential for SM production, SMGCs were predicted using a SMURF-like prediction tool [30] for all species except *N. crassa* and *A. sojae* since these were not sequenced and/or annotated by JGI and thus have not the required generated data and dissimilar gene calling methods.

An overview of the predicted main enzyme responsible for the formation of the core chemical structure of the resulting compound (backbone enzyme) for each species can be seen in Figure 7C. Within the 28 *Aspergillus* species, there is a total of 1,972 predicted SMGCs and for the 22 section *Flavi* species the total is 1606 SMGCs with an average of 73 per species. The maximum of SMGCs is 84 and is found in *A. leporis* while the minimum is 37 for *A. coremiformiis*.

Similar gene clusters (based on sequence similarity of backbone and tailoring enzymes) are expected to code for proteins producing compounds with a similar chemical core structure, therefore, creating families of gene clusters allows us to investigate the SM production potential, evaluate SM diversity and presence/absence patterns across the section. These families of SMGCs were constructed using the method described in Theobald et al. [78, 80] (a table of the predicted clusters and SMGC families can be found in Additional file B3). For the entire dataset with predicted SM 477 cluster families were constructed, where the section *Flavi* SMGCs belong to 308 cluster families. Several cluster families only contain one member of the section *Flavi* species 150 clusters have no 'similar' cluster in the other species in the dataset. The singlet SMGC families are expected to produce compounds specific to one species in the set, or alternatively be incorrectly identified gene clusters or clusters with aberrantly predicted gene models. Families with members in more than one species are expected to be characteristic for a larger group of organisms, as such, 90 SMGC families are found in more than 5 organisms. There are 3 SMGC families found in all the species, 1 family found in all the *Aspergilli* and 2 section *Flavi* specific cluster families. There are also clade specific SMGC families, the *A. flavus* clade shares 1 while *A. tamarii* and *A. nomius* has 0 and 2 families respectively indicating that the *A. flavus* and *A. nomius* clade are closely related groups of species. The number of unique cluster families ranges from 1-3 within the *A. flavus* clade with *A. minisclerotigenes* as an exception with 6 unique families. These numbers are rather low compared with the other species supporting that this clade is very closely related. For species not in the *A. flavus, A. tamarii* or *A. nomius* clade the number of unique cluster families is mainly above 10 showing high diversity and a large potential for novel compounds.

Within section *Nigri* 2,717 SMGCs were identified in 37 species, distributed in 455 SMGC families giving 73.43 clusters per species and 5.97 clusters pr family [80]. Comparing across all the species in this study, the number of clusters per species was 70.42 with 4.13 clusters per family. The number of clusters per species in this study is slightly lower however the number of members per SMGC family is also lower, which indicates a greater diversity in secondary metabolism in section *Flavi* compared to section *Nigri*.

In the *Nigri* study [80] it was shown that 82% of cluster families were found in less than 10 organisms

and 49% were only found in one species with an average of 8.75 unique clusters per species comparing to section *Flavi* the numbers are 88% for cluster families found in less than 10 organisms, 61% found only in one species and 10.31 unique clusters per species. These results further supports that the species in this study are more diverse than in the *Nigri* study. This diversity is mirrored by the number of species included from a section. The more species from a section the closer related they are and the more likely they are to share SMGCs, thus resulting in a lower diversity. Focusing only on the section *Favi* species there are on average 6.8 unique SMGCs per species.

Dereplication of secondary metabolism in section Flavi

To asses the potential for SM production qualitatively, we used a pipeline of dereplication, Figure 7B [78] where predicted clusters are associated with verified known clusters (from the MIBiG database [53]) in a guilt-by-association method. Known gene clusters were coupled to SMGC families making it possible to predict the compounds or derivates thereof a species can potentially produce. Cluster families containing one cluster highly similar to a known compound cluster are labelled after the known compound but further manual inspection of the cluster family is needed in order to evaluate these predictions. Using the dereplication approach 20 cluster families were coupled to a compound family. Clusters similar to the Naphthopyrone [52] cluster were found in all *Flavi* species and so was the nidulanin A [2] cluster family except for *A. leporis*. Clusters similar to Azanigerone [86], 4,4'-piperazine-2,5-diyldimethyl-bisphenol and aflavarin [6]/ endocrocin [5, 41] were identified in all *Flavi* species except *A. coremiiformis*, *A. avenaceus*; *A. avenaceus*; and *A. coremiiformis, A. avenaceus* and *A. leporis* respectively. Clusters similar to Alternariol [1] and aspirochlorine [9] are found in all clade *A. flavus* and *A. tamarii* species plus *A. bertholletius* and *A. avenaceus*. These SMGC families follow the phylogenetic groups with a few exceptions. Clusters similar to the Asperfuranone [10] cluster does not follow the phylogenic groups but is found in *A. tamarii, A. pseudotamarii, A. nidulans* and *A. terreus*. Likewise clusters similar to the pseurotin A [48] or Fumagillin [42] cluster were found in a pattern of species not following the phylogeney (*A. caelatus, A. pseudocaelatus, A. steynii, A. avenaceus, A. leporis* and *A. fumigatus*).

Using the dereplication pipeline we have been able to identify 20 cluster families similar to known biosynthetic clusters. We have thus been able to identify potential producers of known toxins such as aflatoxin and aspirochlorine. Notably, some cluster families mirror the species phylogeny and have representatives in all species within a certain grouping while others are found in multiple species discontinuously across the breadth of the phylogenetic tree possibly indicating different evolutionary paths.

Linking secondary metabolites to clusters by combining chemical and genome analysis

We were interested in linking compounds and clusters based on presence absence pattern of produced compounds and predicted clusters. We therefore created a heatmap of all the cluster families found in at least 5 species, added the predicted compound families from the MIBiG dereplication plus manually curated compound families from a literature survey (Additional files B9). In addition, we created chemical data of the *Flavi* species grown on CYA (Additional file Table B4) to see what compounds are produced by which species and use this to identify the gene cluster responsible for the production.

One compound family, miyakamides, were found to be produced by *A. sojae, A. nomius, parasiticus, A. novoparasiticus* and *A. transmontanensis*. Miyakamides have been isolated from an *A. flavus* species and have been shown to have antibiotic properties [67]. The biosynthetic gene cluster producing miyakamides is not known, but from the chemical structure we performed retro-biosynthesis and predicted that the biosynthetic gene cluster should be contain a NRPS with three adenylation domains (or potentially two since two of the amino acids are highly similar), an N-methyltransferase, an acetyltransferase and potentially a decarboxylase/dehydrogenase, additional file Figure B10A. Holding this information together we searched for cluster families with members in all the miyakamide producing species having NRPS backbones with 3 or 2 adenylation backbones and a methyltransferase domain. With these criteria, only one cluster family met the requirements. The cluster family has has a NRPS backbone with three predicted A domains in *A. transmontanensis* and two predicted in *A. parasiticus, A. novoparasiticus* and *A. nomius*. The NRPS contains a methyltransferase domain, hence two of the required activities are found in one predicted gene. The size of the predicted clusters in these species varies from one to nine genes, the difference are most likely caused by SMGC predictions errors. The predicted functions are also varying. We took sequences for the predicted gene clusters and neighbouring genes and created synteny plots of all the clusters belonging to the cluster family to see which parts are conserved and shared among the species. This conservation gives an indication of which

genes are involved in the biosynthesis, additional Figure B10B. Of the predicted clusters and surrounding genes there are several genes only conserved among some of the species and a small part, containing the NRPS and two small genes with unknown function, widely conserved. These conserved genes are most likely involved in the miyakamide biosynthesis. The clusters of*A. transmontanensi* and *A. parvisclerotigenus* are located at the end of a scaffold, and do not have the two conserved genes in the clusters. In both cases the NRPS is located at the end of a scaffold and homologs of the conserved genes are found the end of another scaffold with high similarity ($\geq$ 94% identity to *A. nomius* genes). Based on the presented data and performed analysis we propose that the identified NRPS along with the two conserved genes of unknown function are involved in the miyakamide biosynthesis.

The aflatoxin biosynthetic gene cluster is highly conserved in the A. flavus, A. nomius and partly in the A. tamarii clade

Perhaps the best known secondary metabolite in section *Flavi* is the highly carcinogenic aflatoxin molecule. Aflatoxins are known to be produced by many section *Flavi* species (*A. arachidicola, A. bombycis, A. flavus, A. minisclerotigenes, A. nomius, A. parvisclerotigenus, A. pseudocaelatus, A. pseudonomius, A. pseudotamarii* and some *A. oryzae* species [79]). We examined the number of species that had the genetic potential to produce aflatoxin, and compared the aflatoxin cluster across the section to assess its variability. From the dereplication analysis, we identified a SMGC family predicted to be involved in sterigmatocystin, aflatoxin or cyclopiazonic-acid, Figure 7B. The first notable observation from the predicted SMGC was that the cluster is found in all the species in the *A. flavus, A. nomius* and *A. tamarii* clade except for *A. tamarii*. A synteny plot of the cluster family (easyfig [74], Additional file Figure B11) shows that the cluster is extremely well conserved with no rearrangements and a high alignment identity for the aflatoxin genes. Only *A. caelatus* seems to have a different version of the cluster where only the *aflB, aflC, aflD* genes are found.

The length of the predicted aflatoxin clusters varies and in *A. parvisclerotigenus, A. arachidicola, A. minisclerotigenes* and *A. sergii* there are several additional genes not known to be involved in aflatoxin biosynthesis included in the clusters. Several of these additional genes have similarity to genes associated with a known cluster responsible for producing cyclopiazonic acid, explaining why two different biosynthetic pathways end up in the same cluster family. Since the clusters are located in very close proximity it is not possible with our prediction method to distinguish and therefore they end up in one large predicted cluster.

The biosynthesis of aflatoxin requires the *aflP* and *aflQ* genes which are responsible for the last steps of the biosynthetic pathway. In most of the predicted clusters these genes were not found. In order to clarify if this was due to cluster prediction errors or if they were not there we used the *A. flavus* genes as query and searched for similar genes using BLASTp. The aflP protein has hits (over 89% identity) for all the species with the predicted aflatoxin cluster, however the query and hit coverage is only around 50%. We made a multiple alignment of the aflP from *A. flavus* and the hits in the other *Flavi* species with the cluster (Additional file Figure B12) to investigate this.

From the alignment it is evident that the *A. flavus* protein has an earlier 5' start site and a gap in the middle compared with the other sequences. We investigated the gene predictions, and the RNA coverage in search of explanations of this (Additional file Figure B13). The RNA coverage (based on transcriptomics) clearly follows the gene predictions and supports the gene prediction in the species sequenced for this study showing that this region is likely to be transcribed in the predicted manner. The *A. flavus* genome was a sequenced previously and annotated with different gene modeling software, which could explain the annotation differences. The genomes in our study were all annotated with gene models supported from next generation sequencing technology RNA-seq giving high quality predictions. The *aflQ* gene has hits with >85% identity and nearly 100% coverage. Lastly the location of the *aflQ* and *aflP* genes were determined relative to the aflatoxin cluster. All *aflQ* genes were 5-10 genes from the predicted aflatoxin gene clusters while the *aflP* genes were 3-8 genes from the predicted aflatoxin clusters. The high similarity and the proximity to the predicted cluster indicate that the genes should be included in the clusters and that all the species do have the *aflP* and *aflQ* genes required required for complete aflatoxin biosynthesis.

As mentioned *A. tamarii* is missing in this cluster family. Since all the other species have the cluster, a possible explanation is a deletion event in *A. tamarii*. To identify a possible genomic locus of the loss, we compared the cluster and surrounding genes from the closest relative *A. pseudotamarii, A. pseudocaelatus* and *A. caelatus* to *A. tamarii*. In *A. pseudotamarii* only the last gene in the predicted cluster is found in *A. tamarii* with high identity (above 90%, marked by blue ring in Additional file Figure B11). In *A. pseudocaelatus* the last and third to last genes have good hits in *A. tamarii* and in *A. caelatus* five of the genes in the right end of the predicted cluster have hits in *A.*

tamarii (marked in Additional file Figure B11). The hits in *A. tamarii* are all found on scaffold 131 close to the end of the scaffold. Comparing the rest of the scaffold 131 from *A. tamarii* with *A. caelatus* the hits are not on the same scaffold as the aflatoxin-like clusters. It indicates that the aflatoxin cluster has undergone similar changes as in *A. caelatus* and even more severe.

Conclusions

We *de novo* sequenced species representing various parts of the *Flavi* section which allowed a section wide comparison illustrating the similarities and diversity within the section. Members of the *Flavi* section has a large genome size compared to other *Aspergilli*. The large genome is reflected in the high number of secondary metabolite gene clusters (SMGCs) and Carbohydrate-Active enZymes (CAZymes) which could be a source of novel compounds and enzymes in the future.

We have shown that the aflatoxin cluster is highly conserved both concerning identity and synteny in the *A. flavus, A. nomius* clade and partly in the *A. tamarii* clade where the cluster is partly lost in *A. caelatus* and completely lost in *A. tamarii*.

The number of species unique proteins is varying but even with the very closely related *A. flavus* clade, most species have above 700 unique proteins illustrating the diversity.

Localization analysis of *A. oryzae* have shown the distribution of unique and SMGC across the genome but with a higher density in the sub-telomeric ends. Synteny analysis have highlighted some tendencies of some highly conserved chromosomes and a few dense non syntenic blocks which could represent rearrangement hot-spots.

Methods

Fungal strains

The species examined in this study were from the IBT Culture Collection of Fungi at the Technical University of Denmark (DTU), unless otherwise noted.

DNA and RNA preparation, sequencing and assembly

The procedure for fungal DNA and RNA preparation, sequencing and assembly has previously been described [33, 80].

Genome annotation

The genomes were annotated using the JGI annotation pipeline [26] as previously described in [33, 80].

Homologous protein families

All predicted proteins from the 31 genomes used in this study were aligned using the BLASTp function from the BLAST+ suite version 2.2.27 with an $e-value \geq 10^{10}$). The 961 whole-genome BLAST tables were analysed to identify homologous proteins and group them into families as describes in [80]. Protein families containing at least one protein from all species were defined as core families while species-unique families were defined as families containing one or more protein(s) from only one species.

Functional annotation

Functional domains were identified in all the proteins using InterPro [18], GO [22] and KOG [76].

Phylogeny

Monocore gene were identified as protein families having exactly one member in each species. Each protein family was aligned using MUSCLE version 3.8.31 (default settings) and then trimmed using gblocks version 0.91b (-t=p -b4=5 -b5=h). Following 200 of these monocore sequences (with length between 150 and 1000 AA) were selected randomly and concatenated and used to construct a phylogentic tree using RaxML version 8.2.8 using PROTGAMMAWAG substitution model and 1000 bootstrapping.

Prediction of Carbohydrate-Active Enzymes

Carbohydrate-Active enZymes (CAZymes) were predicted using the Carbohydrate-Active Enzymes database (CAZy; www.cazy.org, [44]) and the method described by Vesth et al. [80].

Prediction of secondary metabolite gene clusters

Secondary metabolite gene clusters (SMGCs) and SMGC families were predicted based on the SMURF algorithm [30] and the method described in Vesth et al. [80].

Annotation of SMGC families using MIBiG

SMGC families were annotated based on the MIBiG database [53] using the method described in Theobald et al. [78].

Genome synteny analysis

Orthologs were defined as a pair of genes found between two genomes from different species by bidirectional best hits using BLASTP with $E-value \geq 10^{10}$). When two genes within 10 kbp on the 1st genome have corresponding orthologs within 10 kbp on the 2nd genome, the region between the two genes was defined as a syntenic block. The distance between the two genes was calculated by the formula, $|PC1 - PC2| - 1/2(LN1 + LN2)$, where PCn and LNn are nucleotide position of the center and nucleotide length of gene "n" (n = 1 or 2), respectively. -

Analysis of chromosomal localization

For visualization of chromosome, chromosomal location and gene density the R package karyoploteR was used [21].

Secondary metabolite gene cluster synteny analysis and visualization

For visulization of cluster synteny and similarity EasyFig was used [74]. The parameters minimum length and meninmum identity was set to 50 bp and 50% respectively.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

Text for this section . . .

Acknowledgements

Text for this section . . .

Author details

[1]Department of Biotechnology and Bioengineering, Technical University of Denmark, Søltoft Plads 223, 2800 Kongens Lyngby, Denmark. [2]US Department of Energy Joint Genome Institute, YYY, XXX Walnut Creek CA USA. [3]Kikkoman Corporation, YYY, XXX Noda Japan. [4] US Department of Energy Joint BioEnergy Institute, YYY, XXX Emeryville CA USA. [5] Kanazawa Institute of Technology, YYY, XXX Kanazawa Japan. [6] Fungal Physiology, Westerdijk Fungal Biodiversity Institute Fungal Molecular Physiology, Utrecht University, Uppsalalaan 8, 3584 CT Utrecht The Netherlands. [7] Department of Microbiology, Faculty of Agriculture and Forestry, University of Helsinki, Viikinkaari 9, Helsinki Finland. [8] Architecture et Fonction des Macromolécules Biologiques, (CNRS UMR 7257, Aix-Marseille University, , 13288 Marseille France. [9] Environmental Molecular Sciences Division, Earth and Biological Sciences Directorate, Pacific Northwest National Laboratory, , Richland WA USA. [10] Department of Plant and Microbial Biology, University of California, , Berkeley CA USA.

References

1. Ahuja, M., Chiang, Y.M., Chang, S.L., Praseuth, M.B., Entwistle, R., Sanchez, J.F., Lo, H.C., Yeh, H.H., Oakley, B.R., Wang, C.C.. Illuminating the diversity of aromatic polyketide synthases in aspergillus nidulans. Journal of the American Chemical Society 2012;134(19):8212–8221.
2. Andersen, M.R., Nielsen, J.B., Klitgaard, A., Petersen, L.M., Zachariasen, M., Hansen, T.J., Blicher, L.H., Gotfredsen, C.H., Larsen, T.O., Nielsen, K.F., Mortensen, U.H.. Accurate prediction of secondary metabolite gene clusters in filamentous fungi. Proceedings of the National Academy of Sciences 2013;110(1):E99–E107.
3. Barbesgaard, P., Heldt-Hansen, H.P., Diderichsen, B.. On the safety of Aspergillus oryzae: a review. Applied Microbiology and Biotechnology 1992;36:569–572.
4. Benoit, I., Culleton, H., Zhou, M., DiFalco, M., Aguilar-Osorio, G., Battaglia, E., Bouzid, O., J M Brouwer, C.P., O El-Bushari, H.B., Coutinho, P.M., Gruben, B.S., Hildén, K.S., Houbraken, J., Alexis Jiménez Barboza, L., Levasseur, A., Majoor, E., Mäkelä, M.R., Narang, H., Trejo-Aguilar, B., van den Brink, J., VanKuyk, P.A., Wiebenga, A., McKie, V., McCleary, B., Tsang, A., Henrissat, B., de Vries, R.P.. Closely related fungi employ diverse enzymatic strategies to degrade plant biomass. Biotechnology for Biofuels 2015;8:107.
5. Berthier, E., Lim, F.Y., Deng, Q., Guo, C.J., Kontoyiannis, D.P., Wang, C.C., Rindy, J., Beebe, D.J., Huttenlocher, A., Keller, N.P.. Low-Volume Toolbox for the Discovery of Immunosuppressive Fungal Secondary Metabolites. PLoS Pathogens 2013;9(4):e1003289.
6. Cary, J.W., Han, Z., Yin, Y., Lohmar, J.M., Shantappa, S., Harris-Coward, P.Y., Mack, B., Ehrlich, K.C., Wei, Q., Arroyo-Manzanares, N., Uka, V., Vanhaecke, L., Bhatnagar, D., Yu, J., Nierman, W.C., Johns, M.A., Sorensen, D., Shen, H., De Saeger, S., Diana Di Mavungu, J., Calvo, A.M.. Transcriptome analysis of Aspergillus flavus reveals veA-dependent regulation of secondary metabolite gene clusters, including the novel aflavarin cluster. Eukaryotic Cell 2015;14(10):983–997.
7. Castresana, J.. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. Mol Biol Evol 2000;17(4):540–552.
8. Chang, P.K., Ehrlich, K., Fujii, I.. Cyclopiazonic Acid Biosynthesis of Aspergillus flavus and Aspergillus oryzae. Toxins 2009;1:74–99.
9. Chankhamjon, P., Boettger-Schmidt, D., Scherlach, K., Urbansky, B., Lackner, G., Kalb, D., Dahse, H.M., Hoffmeister, D., Hertweck, C.. Biosynthesis of the halogenated mycotoxin aspirochlorine in koji mold involves a cryptic amino acid conversion. Angewandte Chemie - International Edition 2014;53(49):13409–13413.
10. Chiang, Y.M., Szewczyk, E., Davidson, A.D., Keller, N., Oakley, B.R., Wang, C.C.. A gene cluster containing two fungal polyketide synthases encodes the biosynthetic pathway for a polyketide, asperfuranone, in aspergillus nidulans. Journal of the American Chemical Society 2009;131(8):2965–2970.
11. Civas, A., Eberhard, R., Le Dizet, P., Petek, F.. Glycosidases induced in Aspergillus tamarii . Mycelial α-d-galactosidases. Technical Report; 1984.
12. Coutinho, P.M., Andersen, M.R., Kolenova, K., Vankuyk, P.A., Benoit, I., Gruben, B.S., Trejo-Aguilar, B., Visser, H., Van Solingen, P., Pakula, T., Seiboth, B., Battaglia, E., Aguilar-Osorio, G., De Jong, J.F., Ohm, R.A., Aguilar, M., Henrissat, B., Nielsen, J., Stålbrand, H., De Vries, R.P.. Post-genomic insights into the plant polysaccharide degradation potential of Aspergillus nidulans and comparison to Aspergillus niger and Aspergillus oryzae. Fungal Genetics and Biology 2009;46:S161–S169.
13. De Vries, R.P., Riley, R., Wiebenga, A., Aguilar-Osorio, G., Amillis, S., Uchima, C.A., Anderluh, G., Asadollahi, M., Askin, M., Barry, K., Battaglia, E., Bayram, Ö., Benocci, T., Braus-Stromeyer, S.A., Cerqueira, G.C., Chen, F., Chen, W., Choi, C., Clum, A., Corrêa, R.A., Santos, D., De Lima Damásio, R., Diallinas, G., Emri, T., Fekete, E., Flipphi, M., Freyberg, S., Gallo, A., Karaffa, L., Karányi, Z., Kraševec, N., Kuo, A., Kusch, H., Labutti, K., Lagendijk, E.L., Lapidus, A., Levasseur, A., Lindquist, E., Lipzen, A., Logrieco, A.F., Maccabe, A., Molnár, Á.P., Mulé, G., Ngan, C.Y., Orejas, M., Orosz, E., Ouedraogo, J.P., Overkamp, K.M., Ram, A.F.J., Ramón, A., Rauscher, S., Record, E., Riaño-Pachón, D.M., Robert, V., Röhrig, J., Ruller, R., Salamov, A., Salih, N.S., Samson, R.A., Sándor, E., Sanguinetti, M., Schütze, T., Sepčić, K., Shelest, E., Sherlock, G., Sophianopoulou, V., Squina, F.M., Sun, H., Susca, A., Todd, R.B., Tsang, A., Unkles, S.E., Van De Wiele, N., Van Rossen-Uffink, D., Velasco De Castro Oliveira, J., Vesth, T.C., Visser, J., Yu, J.H., Zhou, M., Andersen, M.R., Archer, D.B., Baker, S.E., Benoit, I., Brakhage, A.A., Braus, G.H., Fischer, R., Frisvad, J.C., Goldman, G.H., Houbraken, J., Oakley, B., Pócsi, I., Scazzocchio, C., Seiboth, B., Vankuyk, P.A., Wortman, J., Dyer, P.S., Grigoriev, I.V.. Comparative genomics reveals high biological diversity and specific adaptations in the industrially and medically important fungal genus Aspergillus. Genome Biology 2017;18(1):28.
14. Edgar, R.C.. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 2004;32(5):1792–1797.
15. Eichler, E.E., Sankoff, D.. Structural dynamics of eukaryotic chromosome evolution. Science 2003;301(45):793–797.
16. Fekete, E., De Vries, R.P., Seiboth, B., VanKuyk, P.A., Sándor, E., Fekete, É., Metz, B., Kubicek, C.P., Karaffa, L.. d-galactose uptake is nonfunctional in the conidiospores of Aspergillus niger. FEMS Microbiology Letters 2012;329(2):198–203.
17. Ferreira, G., Boer, C.G., Peralta, R.M.. Production of xylanolytic enzymes by Aspergillus tamarii in solid state fermentation. FEMS Microbiology Letters 1999;173:335–339.
18. Finn, R.D., Attwood, T.K., Babbitt, P.C., Bateman, A., Bork, P., Bridge, A.J., Chang, H.Y., Dosztanyi, Z., El-Gebali, S., Fraser, M., Gough, J., Haft, D., Holliday, G.L., Huang, H., Huang, X., Letunic, I., Lopez, R., Lu, S., Marchler-Bauer, A., Mi, H., Mistry, J., Natale, D.A., Necci, M., Nuka, G., Orengo, C.A., Park, Y.,

Pesseat, S., Piovesan, D., Potter, S.C., Rawlings, N.D., Redaschi, N., Richardson, L., Rivoire, C., Sangrador-Vegas, A., Sigrist, C., Sillitoe, I., Smithers, B., Squizzato, S., Sutton, G., Thanki, N., Thomas, P.D., Tosatto, S.C.E., Wu, C.H., Xenarios, I., Yeh, L.S., Young, S.Y., Mitchell, A.L.. InterPro in 2017—beyond protein family and domain annotations. Nucleic Acids Research 2017;45:D190–D199.

19. Galagan, J.E., Calvo, S.E., Cuomo, C., Ma, L.J., Wortman, J.R., Batzoglou, S., Lee, S.I., Baştürkmen, M., Spevak, C.C., Clutterbuck, J., Kapitonov, V., Jurka, J., Scazzocchio, C., Farman, M., Butler, J., Purcell, S., Harris, S., Braus, G.H., Draht, O., Busch, S., D'Enfert, C., Bouchier, C., Goldman, G.H., Bell-Pedersen, D., Griffiths-Jones, S., Doonan, J.H., Yu, J., Vienken, K., Pain, A., Freitag, M., Selker, E.U., Archer, D.B., Peñalva, M.A., Oakley, B.R., Momany, M., Tanaka, T., Kumagai, T., Asai, K., Machida, M., Nierman, W.C., Denning, D.W., Caddick, M., Hynes, M., Paoletti, M., Fischer, R., Miller, B., Dyer, P., Sachs, M.S., Osmani, S.A., Birren, B.W.. Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae. Nature 2005;438(7071):1105–15.
20. Geiser, D.M., Pitt, J.I., Taylor, J.W.. Cryptic speciation and recombination in the aflatoxin-producing fungus Aspergillus flavus (population genetics mycotoxin biological species microbial evolution). PNAS 1998;95(1):388–393.
21. Gel, B., Serra, E.. KaryoploteR: An R/Bioconductor package to plot customizable genomes displaying arbitrary data. Bioinformatics 2017;33(19):3088–3090.
22. Gene Ontology Consortium, T.. Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Research 2017;45:D331–D338.
23. Gene Ontology Consortium, T., Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.. Gene Ontology: tool for the unification of biology. Nature Genetics 2000;25:25–29.
24. Gibbons, J.G., Salichos, L., Slot, J.C., Rinker, D.C., Mcgary, K.L., King, J.G., Klich, M.A., Tabb, D.L., Mcdonald, W.H., Rokas, A.. The Evolutionary Imprint of Domestication on Genome Variation and Function of the Filamentous Fungus Aspergillus oryzae. Current Biology 2012;22(15):1403–1409.
25. Gourama, H.. Aspergillus flavus and Aspergillus parasiticus: Aflatoxigenic fungi of concern in foods and feeds: A review. Journal of Food Protection 1995;58(12):1395–1404.
26. Grigoriev, I.V., Martinez, D.A., Salamov, A.A.. Fungal genomic annotation. Applied Mycology and Biotechnology 2006;6(C):123–142.
27. Hedayati, M.T., Pasqualotto, A.C., Warn, P.A., Bowyer, P., Denning, D.W.. Aspergillus flavus: human pathogen, allergen and mycotoxin producer. Microbiology 2007;153:1677–1692.
28. Heinen, P.R., Bauermeister, A., Ribeiro, L.F., Messias, J.M., Almeida, P.Z., Moraes, L.A., Vargas-Rechia, C.G., de Oliveira, A.H., Ward, R.J., Filho, E.X., Kadowaki, M.K., Jorge, J.A., Polizeli, M.L.. GH11 xylanase from Aspergillus tamarii Kita: Purification by one-step chromatography and xylooligosaccharides hydrolysis monitored in real-time by mass spectrometry. International Journal of Biological Macromolecules 2018;108:291–299.
29. Ichishima, E.. Development of enzyme technology for Aspergillus oryzae, A. sojae, and A. luchuensis, the national fungi of Japan. Bioscience, Biotechnology and Biochemistry 2016;80(9):1681–1692.
30. Khaldi, N., Seifuddin, F.T., Turner, G., Haft, D., Nierman, W.C., Wolfe, K.H., Fedorova, N.D.. SMURF: Genomic mapping of fungal secondary metabolite clusters. Fungal Genetics and Biology 2010;47:736–741.
31. Kim, J.d.. Production of Xylanolytic Enzyme Complex from Aspergillus flavus using Agri- cultural Wastes. Mycobiology 2005;33(2):84–89.
32. Kimura, I., Sasahara, H., Tajima, S.. Purification and characterization of two xylanases and an arabinofuranosidase from Aspergillus sojae. Journal of Fermentation and Bioengineering 1995;80(4):334–339.
33. Kis-Papo, T., Weig, A.R., Riley, R., Peršoh, D., Salamov, A., Sun, H., Lipzen, A., Wasser, S.P., Rambold, G., Grigoriev, I.V., Nevo, E., Kholodny, M.G.. Genomic adaptations of the halophilic Dead Sea filamentous fungus Eurotium rubrum. Nature communications 2014;5:3745.
34. Klich, M.A.. Aspergillus flavus: the major producer of aflatoxin. Molecular Plant Pathology 2007;8(6):713–722.
35. Klich, M.A., Pitt, J.I.. DIFFERENTIATION OF ASPERGILLUS FLAVUS FROM A. PARASITICUS AND OTHER CLOSELY RELATED SPECIES. Transactions of the British Mycological Society 1988;91(1):99–108.
36. Kobayashi, T., Abe, K., Asai, K., Gomi, K., Juvvadi, P.R., Kato, M., Kitamoto, K., Takeuchi, M., Machida, M.. Genomics of Aspergillus oryzae. Bioscience, Biotechnology, and Biochemistry 2007;71(3):646–670.
37. Kocsubé, S., Perrone, G., Magistà, D., Houbraken, J., Varga, J., Szigeti, G., Hubka, V., Hong, S.B., Frisvad, J.C., Samson, R.A.. Aspergillus is monophyletic: Evidence from multiple gene phylogenies and extrolites profiles. Studies in Mycology 2016;85:199–213.
38. Krishnan, S., Manavathu, E.K., Chandrasekar, P.H.. Aspergillus flavus : an emerging non- fumigatus Aspergillus species of significance. Mycoses 2009;52(3):206–222.
39. Kurtzman, C.P., Smiley, M.J., Robnett, C.J., Wicklow, D.T., Wickl, D.T.. DNA Relatedness among Wild and Domesticated Species in the Aspergillus flavus Group. Mycologia 1986;78(6):955–959.
40. Lebar, M.D., Cary, J.W., Majumdar, R., Carter-Wientjes, C.H., Mack, B.M., Wei, Q., Uka, V., De Saeger, S., Diana Di Mavungu, J.. Identification and functional analysis of the aspergillic acid gene cluster in Aspergillus flavus. Fungal Genetics and Biology 2018;116:14–23.
41. Lim, F.Y., Hou, Y., Chen, Y., Oh, J.H., Lee, I., Bugni, T.S., Keller, N.P.. Genome-based cluster deletion reveals an endocrocin biosynthetic pathway in Aspergillus fumigatus. Applied and Environmental Microbiology 2012;78(12):4117–4125.
42. Lin, H.C., Chooi, Y.H., Dhingra, S., Xu, W., Calvo, A.M., Tang, Y.. The fumagillin biosynthetic gene cluster in Aspergillus fumigatus encodes a cryptic terpene cyclase involved in the formation of β-trans-bergamotene. Journal of the American Chemical Society 2013;135(12):4616–4619.
43. Linz, J.E., Wee, J., Roze, L.V.. Aspergillus parasiticus SU-1 Genome Sequence, Predicted Chromosome Structure, and Comparative Gene Expression under Aflatoxin-Inducing Conditions: Evidence that Differential Expression Contributes to Species Phenotype. Eukaryotic cell 2014;13(8):1113–1123.
44. Lombard, V., Ramulu, H.G., Drula, E., Coutinho, P.M., Henrissat, B.. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Research 2014;42(Database issue):D490–D495.
45. Machida, M., Asai, K., Sano, M., Tanaka, T., Kumagai, T., Terai, G., Kusumoto, K.I., Arima, T., Akita, O., Kashiwagi, Y., Abe, K., Gomi, K., Horiuchi, H., Kitamoto, K., Kobayashi, T., Takeuchi, M., Denning, D.W., Galagan, J.E., Nierman, W.C., Yu, J., Archer, D.B., Bennett, J.W., Bhatnagar, D., Cleveland, T.E., Fedorova, N.D., Gotoh, O., Horikawa, H., Hosoyama, A., Ichinomiya, M., Igarashi, R., Iwashita, K., Juvvadi, P.R., Kato, M., Kato, Y., Kin, T., Kokubun, A., Maeda, H., Maeyama, N., Maruyama, J.I., Nagasaki, H., Nakajima, T., Oda, K., Okada, K., Paulsen, I., Sakamoto, K., Sawano, T., Takahashi, M., Takase, K., Terabayashi, Y., Wortman, J.R., Yamada, O., Yamagata, Y., Anazawa, H., Hata, Y., Koide, Y., Komori, T., Koyama, Y., Minetoki, T., Suharnan, S., Tanaka, A., Isono, K., Kuhara, S., Ogasawara, N., Kikuchi, H., The, . Genome sequencing and analysis of Aspergillus oryzae. Nature 2005;438:1157–1161.
46. Machida, M., Yamada, O., Gomi, K.. Genomics of Aspergillus oryzae: Learning from the History of Koji Mold and Exploration of Its Future. DNA Research 2008;15:173–183.
47. Mahmoud, M., Al-Othman, M., Abd-El-Aziz, A., Metwaly, H., Mohamed, H.. Expression of genes encoding cellulolytic enzymes in some Aspergillus species. Genetics and Molecular Research 2016;15(4):15048913.
48. Maiya, S., Grundmann, A., Li, X., Li, S.M., Turner, G.. Identification of a hybrid PKS/NRPS required for pseurotin A

biosynthesis in the human pathogen Aspergillus fumigatus. ChemBioChem 2007;8(14):1736–1743.
49. Makhuvele, R., Ncube, I., van Rensburg, E.L.J., La Grange, D.C.. Isolation of fungi from dung of wild herbivores for application in bioethanol production. Brazilian Journal of Microbiology 2017;48:648–655.
50. Martinez, D., Berka, R.M., Henrissat, B., Saloheimo, M., Arvas, M., Baker, S.E., Chapman, J., Chertkov, O., Coutinho, P.M., Cullen, D., Danchin, E.G., Grigoriev, I.V., Harris, P., Jackson, M., Kubicek, C.P., Han, C.S., Ho, I., Larrondo, L.F., De Leon, A.L., Magnuson, J.K., Merino, S., Misra, M., Nelson, B., Putnam, N., Robbertse, B., Salamov, A.A., Schmoll, M., Terry, A., Thayer, N., Westerholm-Parvinen, A., Schoch, C.L., Yao, J., Barbote, R., Nelson, M.A., Detter, C., Bruce, D., Kuske, C.R., Xie, G., Richardson, P., Rokhsar, D.S., Lucas, S.M., Rubin, E.M., Dunn-Coleman, N., Ward, M., Brettin, T.S.. Genome sequencing and analysis of the biomass-degrading fungus Trichoderma reesei (syn. Hypocrea jecorina). Nature Biotechnology 2008;26(5):553–560.
51. Mata-Gómez, M.A., Heerd, D., Oyanguren-García, I., Barbero, F., Rito-Palomares, M., Fernández-Lahore, M.. A novel pectin-degrading enzyme complex from Aspergillus sojae ATCC 20235 mutants. Journal of the Science of Food and Agriculture 2015;95:1554–1561.
52. Mayorga, M.E., Timberlake, W.E.. The developmentally regulated Aspergillus nidulans wA gene encodes a polypeptide homologous to polyketide and fatty acid synthases. Molecular and General Genetics 1992;235(2-3):205–212.
53. Medema, M.H., Kottmann, R., Yilmaz, P., Cummings, M., Biggins, J.B., Blin, K., de Bruijn, I., Chooi, Y.H., Claesen, J., Coates, R.C., Cruz-Morales, P., Duddela, S., Düsterhus, S., Edwards, D.J., Fewer, D.P., Garg, N., Geiger, C., Gomez-Escribano, J.P., Greule, A., Hadjithomas, M., Haines, A.S., Helfrich, E.J.N., Hillwig, M.L., Ishida, K., Jones, A.C., Jones, C.S., Jungmann, K., Kegler, C., Kim, H.U., Kötter, P., Krug, D., Masschelein, J., Melnik, A.V., Mantovani, S.M., Monroe, E.A., Moore, M., Moss, N., Nützmann, H.W., Pan, G., Pati, A., Petras, D., Reen, F.J., Rosconi, F., Rui, Z., Tian, Z., Tobias, N.J., Tsunematsu, Y., Wiemann, P., Wyckoff, E., Yan, X., Yim, G., Yu, F., Xie, Y., Aigle, B., Apel, A.K., Balibar, C.J., Balskus, E.P., Barona-Gómez, F., Bechthold, A., Bode, H.B., Borriss, R., Brady, S.F., Brakhage, A.A., Caffrey, P., Cheng, Y.Q., Clardy, J., Cox, R.J., De Mot, R., Donadio, S., Donia, M.S., van der Donk, W.A., Dorrestein, P.C., Doyle, S., Driessen, A.J.M., Ehling-Schulz, M., Entian, K.D., Fischbach, M.A., Gerwick, L., Gerwick, W.H., Gross, H., Gust, B., Hertweck, C., Höfte, M., Jensen, S.E., Ju, J., Katz, L., Kaysser, L., Klassen, J.L., Keller, N.P., Kormanec, J., Kuipers, O.P., Kuzuyama, T., Kyrpides, N.C., Kwon, H.J., Lautru, S., Lavigne, R., Lee, C.Y., Linquan, B., Liu, X., Liu, W., Luzhetskyy, A., Mahmud, T., Mast, Y., Méndez, C., Metsä-Ketelä, M., Micklefield, J., Mitchell, D.A., Moore, B.S., Moreira, L.M., Müller, R., Neilan, B.A., Nett, M., Nielsen, J., O'Gara, F., Oikawa, H., Osbourn, A., Osburne, M.S., Ostash, B., Payne, S.M., Pernodet, J.L., Petricek, M., Piel, J., Ploux, O., Raaijmakers, J.M., Salas, J.A., Schmitt, E.K., Scott, B., Seipke, R.F., Shen, B., Sherman, D.H., Sivonen, K., Smanski, M.J., Sosio, M., Stegmann, E., Süssmuth, R.D., Tahlan, K., Thomas, C.M., Tang, Y., Truman, A.W., Viaud, M., Walton, J.D., Walsh, C.T., Weber, T., van Wezel, G.P., Wilkinson, B., Willey, J.M., Wohlleben, W., Wright, G.D., Ziemert, N., Zhang, C., Zotchev, S.B., Breitling, R., Takano, E., Glöckner, F.O.. The Minimum Information about a Biosynthetic Gene cluster. Nature Chemical Biology 2015;11(9):625–631.
54. Mellon, J.E., Cotty, P.J., Callicott, K.A., Abbas, H.. Identification of a Major Xylanase from Aspergillus flavus as a 14-kD Protein. Mycopathologia 2011;172:299–305.
55. Moore, G.G., Mack, B.M., Beltz, S.B.. Genomic sequence of the aflatoxigenic filamentous fungus Aspergillus nomius. BMC Genomics 2015;16:551.
56. Moore, G.G., Mack, B.M., Beltz, S.B., Gilbert, M.K.. Draft Genome Sequence of an Aflatoxigenic Aspergillus Species, A. bombycis. Genome Biology and Evolution 2016;8(11):3297–3300.
57. Moore, G.G., Mack, B.M., Beltz, S.B., Puel, O.. Genome sequence of an aflatoxigenic pathogen of Argentinian peanut, Aspergillus arachidicola. BMC Genomics 2018;19:189.
58. Moreira, F.G., Lenartovicz, V., Peralta, R.M.. A thermostable maltose-tolerant α-amylase from Aspergillus tamarii. Journal of Basic Microbiology 2004;44:29–35.
59. Mäkelä MR DiFalco M, M.E.N.T.W.A.H.K.P.M.G.I.T.A.d.V.R.. Genomic and exoproteomic diversity in plant biomass degradation approaches among Aspergilli 2018;xx(X):x–xx.
60. Nicholson, M.J., Koulman, A., Monahan, B.J., Pritchard, B.L., Payne, G.A., Scott, B.. Identification of two aflatrem biosynthesis gene loci in Aspergillus flavus and metabolic engineering of Penicillium paxilli to elucidate their function. Applied and Environmental Microbiology 2009;75(23):7469–7481.
61. Nierman, W.C., May, G., Kim, H.S., Anderson, M.J., Chen, D., Denning, D.W.. What the Aspergillus genomes have told us. Medical Mycology Supplement 2005;43:S3–S5.
62. Nierman, W.C., Yu, J., Fedorova-Abrams, N.D., Losada, L., Cleveland, T.E., Bhatnagar, D., Bennett, J.W., Dean, R., Payne, G.A.. Genome Sequence of Aspergillus flavus NRRL 3357, a Strain That Causes Aflatoxin Contamination of Food and Feed. Genome Announcements 2015;3(2):e00168–15.
63. Paradis, E., Claude, J., Strimmer, K.. APE: Analyses of Phylogenetics and Evolution in R language. BIOINFORMATICS APPLICATIONS NOTE 2004;20(2):289–290.
64. Saroj, P., P, M., Narasimhulu, K.. Characterization of thermophilic fungi producing extracellular lignocellulolytic enzymes for lignocellulosic hydrolysis under solid-state fermentation. Bioresources and Bioprocessing 2018;5:31.
65. Sato, A., Oshima, K., Noguchi, H., Ogawa, M., Takahashi, T., Oguma, T., Koyama, Y., Itoh, T., Hattori, M., Hanya, Y.. Draft Genome Sequencing and Comparative Analysis of Aspergillus sojae NBRC4239. DNA Research 2011;18:165–176.
66. Sen, S., Ray, L., Chattopadhyay, P.. Production, purification, immobilization, and characterization of a thermostable β-galactosidase from aspergillus alliaceus. Applied Biochemistry and Biotechnology 2012;167:1938–1953.
67. Shiomi, K., Hatae, K., Yuichi Yamaguchi, Õ., Mas Hiroshi Tomoda, R., Kobayashi, S., Omura, S.. New Antibiotics Miyakamides Produced by a Fungus. JOURNAL OF ANTIBIOTICS 2002;55(11):952–961.
68. da Silva, A.C., Soares de França Queiroz, A.E., Correiab, P.C., Brandão-costaa, R., Moreira, K.A., Porto, A.L.F., de Medeiros, E.V.. Production and Characterization of Xylanase from Aspergillus parasiticus URM 5963 Isolated from Soil of Caatinga. Technical Report 2; 2016.
69. Simã, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M.. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 2015;31(19):3210–3212.
70. de Siqueira, F.G., de Siqueira, A.G., de Siqueira, E.G., Carvalho, M.A., Peretti, B.M.P., Jaramillo, P.M.D., Teixeira, R.S.S., Dias, E.S., Félix, C.R., Filho, E.X.F.. Evaluation of holocellulase production by plant-degrading fungi grown on agro-industrial residues. Biodegradation 2010;21(5):815–824.
71. de Souza, C.G.M., Girardo, N.S., Costa, M.A.F., Peralta, R.M.. Influence of growth conditions on the production of xylanolytic enzymes byAspergillus flavus. Journal of Basic Microbiology 1999;39(3):155–160.
72. de Souza, D.F., de Souza, C.G.M., Peralta, R.M.. Effect of easily metabolizable sugars in the production of xylanase by Aspergillus tamarii in solid-state fermentation. Process Biochemistry 2001;36:835–838.
73. Stamatakis, A.. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 2014;30(9):1312–1313.
74. Sullivan, M.J., Petty, N.K., Beatson, S.A.. Easyfig: a genome comparison visualizer. BIOINFORMATICS APPLICATIONS NOTE 2011;27(7):1009–101010.
75. Taniwaki, M.H., Pitt, J.I., Iamanaka, B.T., Sartori, D., Copetti, M.V., Balajee, A., Helena, M., Fungaro, P., Frisvad, J.C.. Aspergillus bertholletius sp. nov. from Brazil Nuts. PLoS ONE 2012;7(8):e42480.

76. Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Smirnov, S., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J., Natale, D.A.. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 2003;4:41.
77. Terada, M., Hayashi, K., Mizunuma, T.. Distinction between Aspergillus oryzae and Aspergillus sojae by the productivity of some hydrolytic enzymes. Nihon shoyu kenkyusho zasshi 1980;6:75–81.
78. Theobald, S., Vesth, T.C., mmer Rendsvig, J.K., Nielsen, K.F., Riley, R., de Abreu, L.M., Salamov, A., Frisvad, J.C., Larsen, T.O., Andersen, M.R., Hoof, J.B.. Uncovering secondary metabolite evolution and biosynthesis using gene cluster networks and genetic dereplication. Manuscript in review ;.
79. Varga, J., Frisvad, J., Samson, R.. Two new aflatoxin producing species, and an overview of Aspergillus section Flavi. Studies in Mycology 2011;69:57–80.
80. Vesth, T.C., Nybo, J.L., Theobald, S., Frisvad, J.C., Larsen, T.O., Nielsen, K.F., Hoof, J.B., Brandl, J., Salamov, A., Riley, R., Gladden, J.M., Phatale, P., Nielsen, M.T., Lyhne, E.K., Kogle, M.E., Strasser, K., McDonnell, E., Barry, K., Clum, A., Chen, C., LaButti, K., Haridas, S., Nolan, M., Sandor, L., Kuo, A., Lipzen, A., Hainaut, M., Drula, E., Tsang, A., Magnuson, J.K., Henrissat, B., Wiebenga, A., Simmons, B.A., Mäkelä, M.R., de Vries, R.P., Grigoriev, I.V., Mortensen, U.H., Baker, S.E., Andersen, M.R.. Investigation of inter- and intra-species variation through genome sequencing of Aspergillus section Nigri. Nature Genetics 2018;Accepted -.
81. Watanabe, K.. Effective use of heterologous hosts for characterization of biosynthetic enzymes allows production of natural products and promotes new natural product discovery. Technical Report 12; 2014.
82. Wollenberg, R.D., Saei, W., Westphal, K.R., Klitgaard, C.S., Nielsen, K.L., Lysøe, E., Gardiner, D.M., Wimmer, R., Sondergaard, T.E., Sørensen, J.L.. Chrysogine Biosynthesis Is Mediated by a Two-Module Nonribosomal Peptide Synthetase. Journal of Natural Products 2017;80(7):2131–2135.
83. Yu, J., Nierman, W.C., Fedorova, N.D., Bhatnagar, D., Cleveland, T.E., Bennett, J.W.. What can the Aspergillus flavus genome offer to mycotoxin research? Mycology 2011;2(3):218–236.
84. Yuan, G.F., Liu, C.S., Chen, C.C.. Differentiation of Aspergillus parasiticus from Aspergillus sojae by Random Amplification of Polymorphic DNA. Applied and Environmental Microbiology 1995;61(6):2384–2387.
85. Yun, C.S., Motoyama, T., Osada, H.. Biosynthesis of the mycotoxin tenuazonic acid by a fungal NRPS-PKS hybrid enzyme. Nature Communications 2015;6:8758.
86. Zabala, A.O., Xu, W., Chooi, Y.H., Tang, Y.. Characterization of a silent azaphilone gene cluster from Aspergillus niger ATCC 1015 reveals a hydroxylation-mediated pyran-ring formation. Chemistry and Biology 2012;19(8):1049–1059.

Tables

Table 1 Percentage of genome syntenicly conserved. .

Species	# Syntenic genes	% of *A. oryzae*
A. parasiticus	8199	68.15%
A. transmontanensis	8238	68.48%
A. arachidicola	8817	73.29%
A. novoparasiticus	8102	67.35%
A. sergii	8091	67.26%
A. flavus	8686	72.20%
A. parvisclerotigenus	9094	75.59%
A. oryzae	–	–
A. minisclerotigenes	8498	70.64%
A. caelatus	7411	61.60%
A. pseudocaelatus	7503	62.37%
A. pseudotamarii	7494	62.29%
A. tamarii	7471	62.10%
A. pseudonomius	7179	59.68%
A. nomius	7269	60.42%
A. bombycis	7863	65.36%
A. bertholletius	6801	56.53%
A. alliaceus	6021	50.05%
A. albertensis	5998	49.86%
A. coremiiformis	5425	45.10%
A. leporis	5800	48.21%
A. avenaceus	5351	44.48%
A. nidulans	4272	35.51%
A. fumigatus	4876	40.53%

Figures

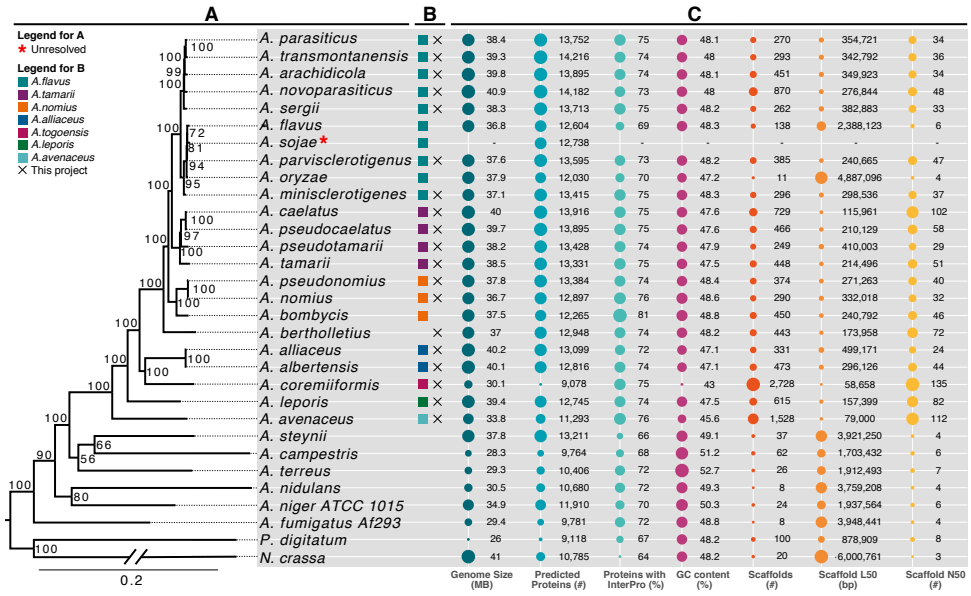


Figure 1 Phylogeny and genome statitics of section Flavi plus 8 other Aspergillus, Penecillium and Neurospora species. A) Phylogenetic tree constructed using RAxML [73], MUSCLE [14], and Gblock [7] based on 200 monocore genes. The red star indicates uncertain/unresolved branches. B) The colors illustrates the clades found within section *Flavi* and the cross indicates species sequenced in this study. C) Seven bubble plots illustrating key genome numbers and sequencing quality parameter. The bubble sizes have been scaled to each panel and are not comparable across panels.

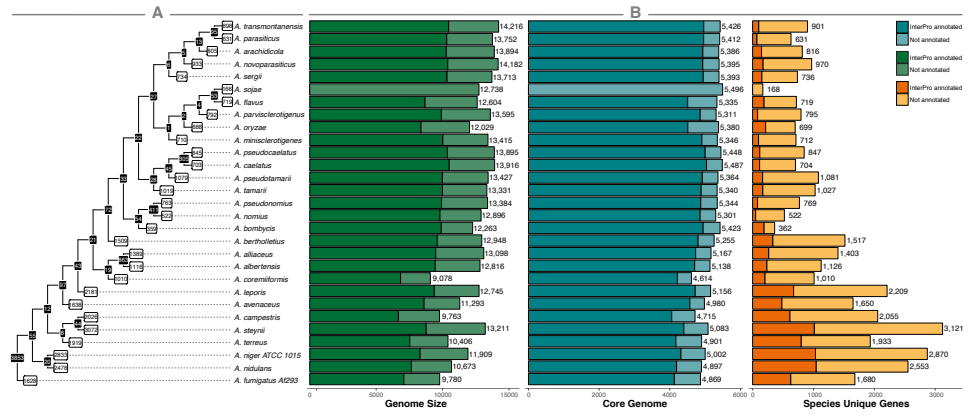


Figure 2 Core, section, clade specific and species unique genes A) A dendrogram representing the phylogenetic relationship between the 31 Aspergilli. The black boxes in the nodes represent the homologous protein families shared among the species branching from that node. The white boxes at the tips represent the protein families unique to that individual species. B) A barplot showing the number of total (green), core (turquoise), and species-unique (orange) proteins for each species. The dark shading illustrates the number of proteins with a least one functional annotation based on InterPro [18].

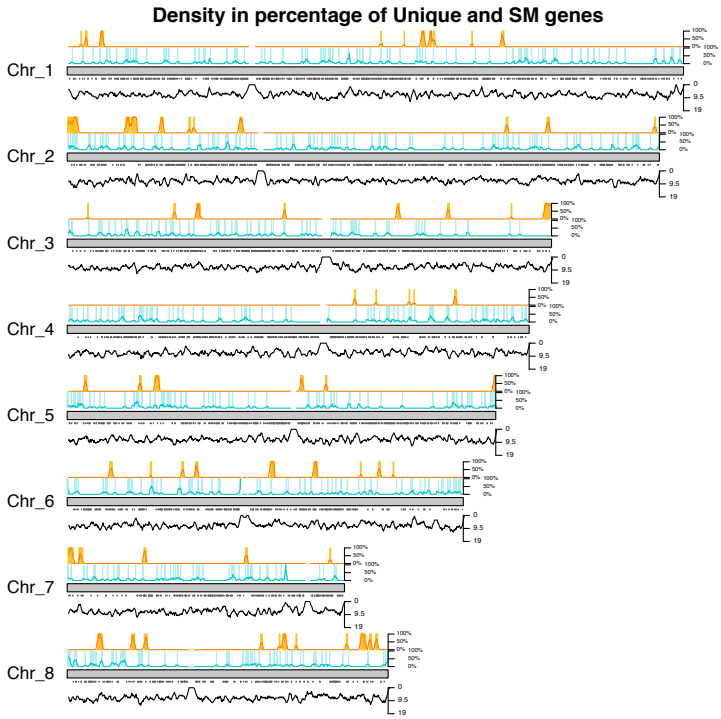


Figure 3 Location of species unique and secondary metabolite genes in the A. oryzae genome The grey bars represent the *A. oryzae* genome. Above the chromosome the unique (turquoise) and secondary metabolite genes (orange) are mapped to the genome, each line represents a gene. The curve shows the percentage of the density calculated from the total number of genes within 30 kbp in steps of 5 kb. Below the genome the core genes are mapped by the grey dots and the density of the total number of genes are shown by the black graph (with a window of 30 kbp).

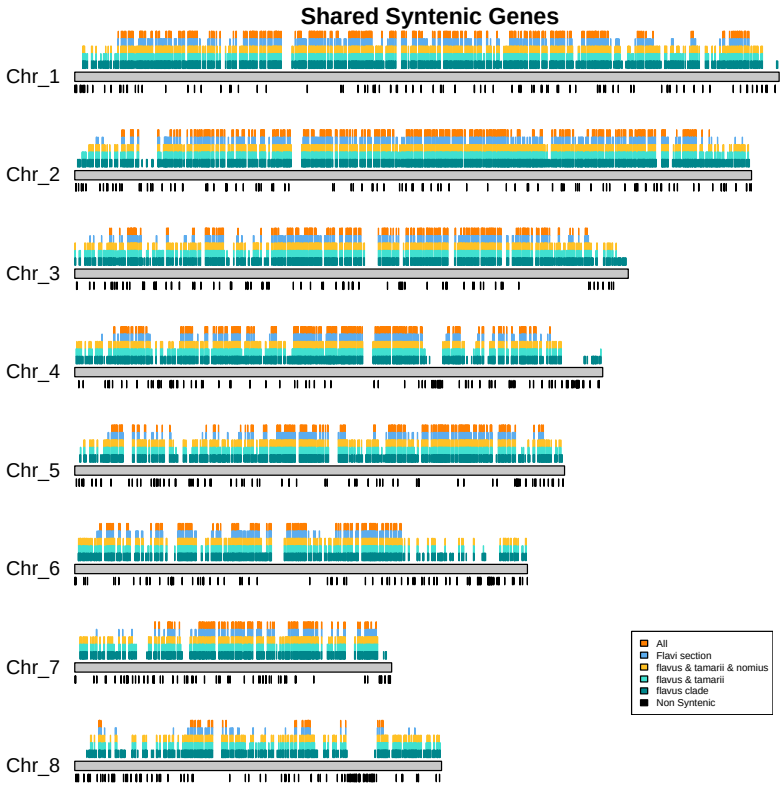


Figure 4 Syntenic genes mapped to the A. oryzae genome The grey bars represent the *A. oryzae* genome. Syntenic genes were found using *A. oryzae* as reference comparing all the other species to this genome. Above the chromosomes syntenic genes shared with all species in a certain group is shown. Dark turquoise - *A. flavus* clade; light turquoise - *A. flavus* and *A. tamarii* clade; yellow - *A. flavus, A. tamarii* and *A. nomius* clade; blue - *Flavi* section; orange - *Flavi* section, *A. nidulans* and *A. fumigatus*. Below the chromosome the non-syntenic genes, found to be non syntenic in all the species, are shown in black.

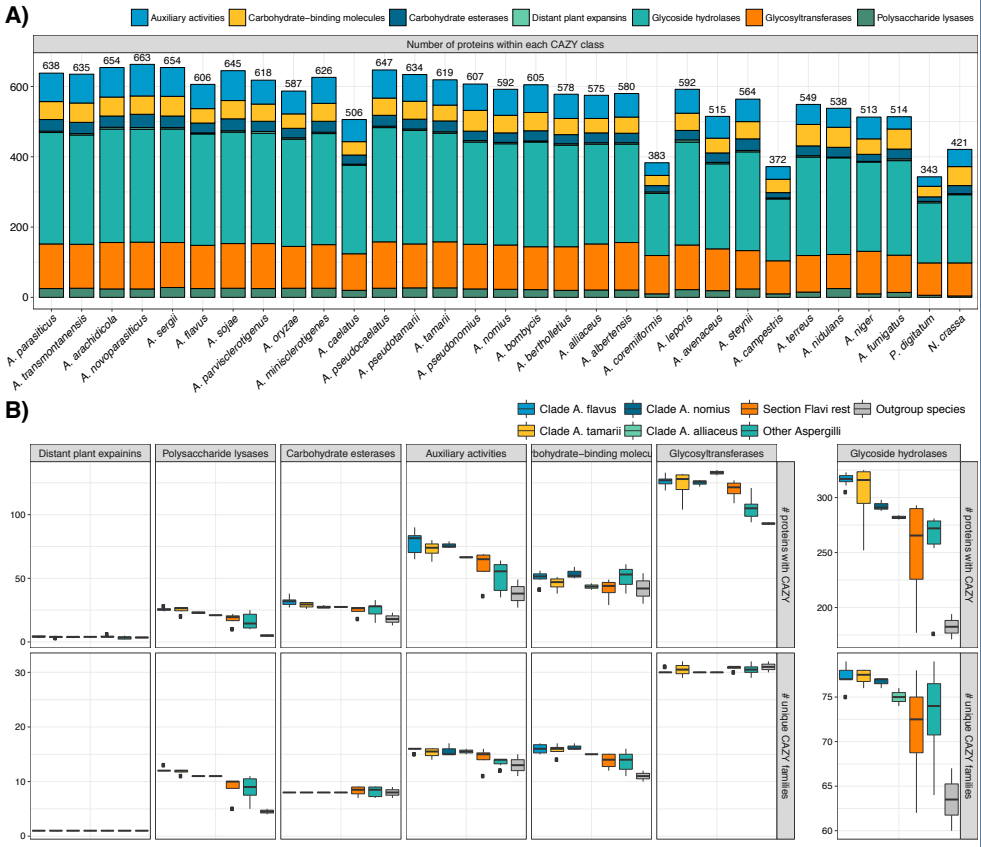


Figure 5 Carbohydrate active enzymes in section Flavi A) The total number of Carbohydrate-Active enZymes (CAZymes) in each species distributed on 7 categories of enzyme activity; auxiliary activities, carbohydrate-binding molecules, carbohydrate esterases, distant plant expansins, glycoside hydrolases, glycosyltransferases, and polysaccharide lysases. B) Boxplot representing the diversity of CAZymes amount and content among clade *A. flavus* (light blue), *A. tamraii* (yellow), *A. nomius* (dark blue), *A. alliaceus* (light turquoise), the rest of the *Flavi* section (orange), other Aspergilli (dark turquoise), non-Aspergillus species (grey). For each CAZyme class the total number of CAZymes (top row) and the number of unique CAZyme families (bottom row) are displayed.

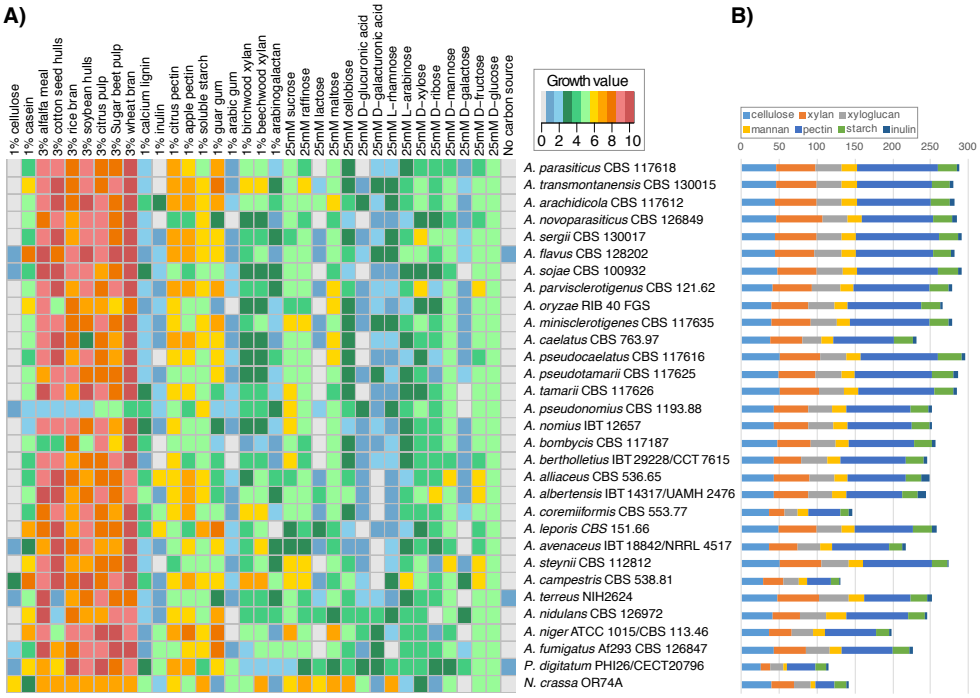


Figure 6 Carbohydrate active enzymes in section Flavi A) Heatmap representing the growth profiles of 23 *Flavi* species and 8 addition species on 35 different media. B) Comparison of the CAZyme sets related to plant biomass degradation in the genomes of species from *Aspergillus* section *Flavi*, and some other fungi. The colours reflect the polysaccharides the enzymes are active on. The blue line indicates the species of section *Flavi*.

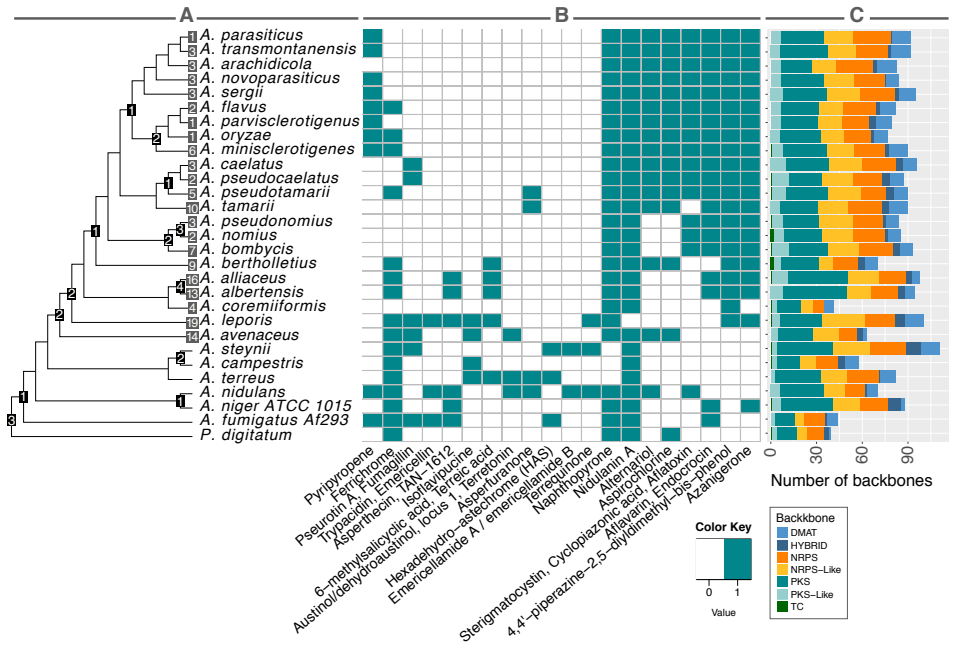


Figure 7 Dereplikation of known compound and predicted secondary metabolite gene backbone genes per species. A) A dendrogram representing the phylogenetic relationship between the species. The black boxes in the nodes represent the secondary metabolite gene cluster (SMGC) families shared among the species branching from that node. If there is no black box there are zero clusters shared. The grey boxes at the tips shows the number of unique SMGC families only found in one species. B) Matrix indicating the presence and absence of SMGC families coupled to known clusters from the MIBiG database [53] for each species. C) Predicted secondary metabolite genes for each species divided by the backbone enzyme. DMAT: Dimethylallyltransferase (Prenyl transferases), HYBRID: A backbone gene containing domains from NRPS and PKS backbones, NRPS: Non-ribosomal peptide synthetase, NRPS-like: Non-ribosomal peptide synthetase like, containing at least two NRPS specific domains and another domain or one NRPS A domain in combination with NAD binding 4 domain or short chain dehydrogenase, PKS: Polyketide synthase, PKS-like: Polyketide synthase like, containing at least two PKS speci
c domains and another domain, TC: Terpene cyclase

Additional Files

Table B1 Growth section *Flavi* quantitatively. Growth analysis of 23 *Flavi* species plus 8 additional species on 35 different growth media, quantitated by growth from 0-10, normalized based on growth on 1% glucose

Table B2 CAZy content in section *Flavi*. Overview of the CAZy content and plant degradation related CAZy content

Table B3 Secondary metabolite gene clusters section *Flavi*. Long format table with all the predicted clusters in the species and the cluster family they belong to. Column 1 - Species, column 2 - JGI protein id of the predicted backbone, column 3 - cluster family number.

Table B4 Compounds produced by *Flavi* species after growth on CYA 7 days.

Figure B1 Phylogenetic tree based on 300 monocore genes. Phylogenetic tree constructed using RAxML [73], MUSCLE [14], and Gblock [7] based on 300 monocore genes.

Figure B2 Phylogenetic tree based on 500 monocore genes. Phylogenetic tree constructed using RAxML [73], MUSCLE [14], and Gblock [7] based on 200 monocore genes.

Figure B3 Single genes phylogenetic trees of 18 monocore families. Phylogenetic trees constructed using RAxML [73] and MUSCLE [14] of 18 single gene tress from monocore families.

Figure B4 Most common InterPro domains in species unique proteins. Bar plot showing the number of species unique proteins with an InterPro domain per species, shown for the most common InterPro annotations [18]. Light grey star indicates p-values below 0.005 and dark grey star indicates p-value below 0.00001 of enrichment in the species unique genes for the specific functional domain.

Figure B5 Most common GO domains in species unique proteins. Bar plot showing the number of species unique proteins with a GO domain per species, shown for the most common GO annotations [22]. Dark grey star indicates p-value below 0.00001 of enrichment in the species unique genes for the specific functional domain.

Figure B6 Most common KOG domains in species unique proteins. Bar plot showing the number of species unique proteins with an KOG domain per species, shown for the most common KOG annotations [76]. P-values indicates enrichment of a certain KOG term in the species unique genes for the specific functional domain. The p-values are indicated by stars of light to dark grey the lightest star indicates p-values below 0.05, followed by p-values below 0.005 and the darkest stars indicates p-value below 0.00001.

Figure B7 Growth section Flavi. Growth analysis of 23 *Flavi* species plus 8 additional species on 35 different growth media.

Figure B8 Phylogenetic tree of GH28. Phylogenetic tree of all proteins assigned to the GH28 CAZy catagory. The GH28 family consits of polygalacturonase. Alignment of the members of GH28 CAZy family found in all section *Flavi* species was created using clustalo. The ML phylogenetic tree was created using the ape package in R [63].

Figure B9 Heatmap cluster families. Cluster families with members in at least 5 species are illustrated by a heatmap. The top rows indicate the backbone enzymes found within the cluster family. Compounds with similar clusters are added from the dereplication using MIBiG marked by orange boxes and black text in addition to manually curated compounds (marked by black boxes and blue text). Aspergillic acid [40], aflatrem [60], chrysogine [82], tenuazonic acid [85], ditryptophenaline [81].

Figure B10 Miyakamide and putative clusters. A) Miyakamide B1 showing the three amino acid parts (organge line), the acetylation (purple circle), the decarboxylation (green circle) and the N-methylation (blue circle). B) Syntenic plot of the putative miyakamide cluster family plus surrounding genes. The syntney plot was generated using EasyFig [74] with the minimum length 50 bp and the minimum identity to 50%. The genes potentially involved in miyakamide production are marked by orange.

Figure B11 Synteny of aflatoxin cluster family. Syntenic plot of the predicted clusters belonging to the 'sterigmatocystin - aflatoxin - cyclopiazonic-acid' cluster family. The syntney plot was generated using EasyFig [74] with the minimum length 50 bp and the minimum identity to 50%. The genes are color coded based on known pathways; aflatoxin (orange), cyclopiazonic-acid (turquoise) and unknown (blue).

Figure B12 Investigation of aflP and aflQ. The *aflP* gene is important for the last steps in the biosynthesis of aflatoxin, but it is missing in most of the predicted aflatoxin clusters. Here is an alignment of the aflP protein against the best hits in the other *Flavi* species.

Figure B13 Geneprediction of aflP. An overview of the gene prediction in the *aflP* gene in *A. paraciticus* and the RNA coverage from JGI at Mycocosm.

4 Resistance gene-directed genome mining

There are thousands of predicted secondary metabolite gene clusters and it takes a significant effort to identify the compounds and test them for bioactivity etc. The idea behind Fungal ResIstance Gene-directed Genome mining (FRIGG) is that we can select for bioactive clusters and knowing the Resistance mechanism we can have an idea of the possible uses of the compound such as fungicides, cholesterol lowering and so on.

In this chapter we will first introduce a pipeline we have developed for identifying putative resistance genes and bioactive biosynthetic gene clusters (section 4.1) followed by experimental investigation of an identified cluster (section 4.2).

4.1 Manuscript III - Pipeline for resistance gene-directed genome mining

We have developed a pipeline for Fungal ResIstance Gene-directed Genome mining (FRIGG) based on the hypothesis of a duplicated essential gene functioning as a resistance gene.

As mentioned in the background (Section 2.3), there are many self-resistance mechanisms. In this pipeline we are targeting a very specific self-resistance mechanism — the duplication of an essential target gene which is inhibited by a secondary metabolite, but the duplicated resistance gene is not and can thus still function. The duplicated resistance gene is found with the biosynthetic gene cluster producing the secondary metabolite. This knowledge is used to identify resistance genes and bioactive clusters and it allows for the identification of novel drug targets.

Whole-genome sequencing and comparative genomics are both prerequisites for the development of this pipeline and will briefly be introduced here. Whole genome sequencing of filamentous fungi have reveal an enormous biological diversity, which was evident from the first whole genomes [25, 23, 24] and have been confirmed with sequencing projects such as the 1000 fungal genomes (http://1000.fungalgenomes.org) and *Aspergillus* sequencing project [92, 93].

Comparative genomics paves the way of finding new insights and patterns from whole genome sequences and have been developed and used extensively in bacteria for many years [94, 95] and these methods are now becoming more extensively used for fungi [96].

Comparative genomics can be used to identify similarities across many species such as orthologous proteins and differences such as specific traits only found in a few species. It is also extensively used to annotate newly sequenced genomes based on knowledge from well studied genomes and from sequence similarity identified using tools such as Basic Local Alignment Search Tool (BLAST) [97]. Other tools have been developed inferring the functional domains from the sequences such as Gene Ontology (GO) [98, 99], eukaryotic orthologous groups (KOG) [100] and InterPro [101, 102].

When the first fungal genomes were sequenced and revealed a large number of putative secondary metabolite gene clusters, combined with comparative genomic tools, the idea of genome mining simply emerged — predicting and identifying natural products based on the genetic information. Genome mining has been extensively used in fungal genomes to identify secondary metabolite gene clusters using sequence similarity of synthase/synthetase genes and predictive tools such as SMURF [66] and antiSMASH (including the fungal version https://fungismash.secondarymetabolites.org/#!/start) [103].

Today the challenge is not only to identify the biosynthetic clusters but to select the most promising clusters and prioritize the experimental effort. We wanted to create a method overcoming the challenge of data overload by exploiting the biological diversity, wealth of whole genome sequences and the opportunities created by comparative genomics. To do this, we created a pipeline identifying biosynthetic gene clusters containing putative self-resistance genes. The developed pipeline is based on comparative genomics, which is extreme powerful, when one has many genomes available. Orthologs are identified in the species based on BLASTp and secondary metabolite gene clusters are predicted using SMURF [66] and this information is coupled and then filtered based in a hypothesis driven approach. The method is described in detail the following manuscript (Manuscript III).

Manuscript III will be submitted to the journal: Fungal Biology and Biotechnology. The additional files to the manuscript can be found in Appendix C. The scripts and data can be found at the Github page: https://github.com/ingek-1/FRIGG_pipeline and this repository: https://files.dtu.dk/u/ox6SPjaekiyxHpI8/Data_tables_FRIGG?l.

RESEARCH

Resistance Gene-Directed Genome Mining of 50 Aspergillus species

Inge Kjærbølling*, Tammi Vesth and Mikael R. Andersen

*Correspondence:
ingek@bio.dtu.dk
Technical University of Denmark, Søltoft Plads 223, 2800 Kongens Lyngby, Denmark
Full list of author information is available at the end of the article

Abstract

Background: Fungal secondary metabolites are a rich source of valuable natural products. Genome sequencing have revealed an enormous potential from predicted biosynthetic gene clusters. It is however currently a time consuming task and an unfeasible task to characterize all biosynthetic gene cluster and to identify possible uses of the compounds. A rational approach is needed to identify promising gene clusters responsible for producing valuable compounds. Several valuable bioactive clusters have been shown to include a resistance gene which is a paralog of the target gene inhibited by the compound. This mechanism can be used to design a rational approach selecting those clusters.

Results: We have developed a pipeline FRIGG (Fungal ReslStance Gene-directed Genome mining) identifying putative resistance genes found in biosynthetic gene clusters based on homology patterns of the cluster genes. The FRIGG pipeline has been run using 51 *Aspergillus* and *Penicillium* genomes, identifying 72 unique protein families with putative resistance genes using various settings in the pipeline. The pipeline was also able to identify the characterized resistance gene *inpE* from the Fellutamide B cluster thereby validating the approach.

Conclusion: We have successfully developed an approach identifying putative valuable bio-active clusters based on a specific resistance mechanism. This approach will be highly useful as an ever increasing amount of genomic data becomes available — the art of identifying and selecting clusters producing novel valuable compounds will only become more crucial.

Keywords: secondary metabolism; resistance; genome mining; Aspergillus

Background

Fungal secondary metabolites are a rich source of bio-active compounds including important pharmaceuticals such as penicillin, cyclosporin and statin [1]. When the first fungal genomes were sequenced, it became clear that the genomes harbour a higher number of secondary metabolite gene clusters than the number of characterized secondary metabolites thus revealing a much larger potential [2, 3, 4, 1]. The number of sequenced genomes is ever increasing mainly due to large sequencing efforts such as the 1000 Fungal Genomes Project of the Department of Energy Joint Genomes Initiative (http://1000.fungalgenomes.org/home/) and the 300 Aspergillus genome project [5, 6] and therefore the number of predicted secondary metabolite gene clusters is steadily increasing.

Despite progress in molecular tools and methods for characterization of secondary metabolite gene clusters, it is still a time-consuming task, making it unfeasible to investigate all predicted secondary metabolite gene clusters. Therefore only a small

fraction of the predicted clusters are characterized and investigated experimentally. With the plethora of predicted secondary metabolite gene clusters (clusters) and the aim of discovering novel bio-active compounds useful as drugs, the question emerges: How do we select the most interesting predicted clusters producing potential valuable drugs such as anti-fungicides, anti-cancer drugs and anti-microbial compounds? To meet this need we have created a pipeline FRIGG (Fungal ResIstance Gene-directed Genome mining) identifying clusters producing likely bio-active compounds based on resistance genes.

Many bio-active compounds are toxic compounds also impairing the organism that synthesize them by inhibiting essential functions, therefore a self-resistance mechanism is needed in order to survive [7, 8, 9]. One known self-resistance mechanism is the duplication of the target gene, where the second version is resistant towards the compound and this second resistant version is most often found as part of the biosynthetic gene cluster producing the toxic compound. This mechanism has been seen in several bacterial instances such as novobiocin [10] and pentalenolactone [11, 12].

More recently this resistance mechanism has also been identified in fungi. Mycophenolic acid (MPA) is produced by *Penicillium brevicompactum* and it inhibits inosine-5'- monophosphate dehydrogenase (IMPDH) which is the rate limiting step in guanine synthesis. The biosynthetic cluster of MPA revealed an additional copy of IMPDH which is insensitive to MPA thus conferring resistance, Figure 1A [13, 14, 15]. Another example is Fellutamide B produced by *A. nidulans* which is a proteasome inhibitor. Within the biosynthetic gene cluster a gene, *inpE*, encoding a proteasome subunit is located and it was shown that this gene confers resistance, Figure 1B [16].

An illustration of the general mechanism can be seen on Figure 1C, where two versions of an enzyme is present and one version is affected by the secondary metabolite (the target) whereas the other version is slightly different, but still with the same function, is not inhibited by the secondary metabolite (the resistance gene). Even though only a few examples of this mechanism have been identified and verified in filamentous fungi so far [13, 16, 17, 18], it is possible that this resistance mechanism is more widely distributed. We have therefore developed a Fungal ResIstance Gene-directed Genome mining (FRIGG) pipeline identifying putative bio-active clusters with resistance genes. The aim of the pipeline is to identify bio-active clusters in a targeted manner thus providing a way of selecting the most interesting predicted clusters producing potential valuable drugs from whole genome sequences.

The immediate advantage of the FRIGG pipeline is that highly likely bioactive clusters are identified. Another major advantage of using this approach is that the target of the compound is inherently known and hence the mode of action. Knowing the target saves a lot of time since linking the compound to the target is extreme difficult and time consuming and several regular drug discovery steps can be eliminated since possible uses of the compound are known from the beginning.

Results

Pipeline set-up and Input data

We were interested in creating a pipeline identifying secondary metabolite gene clusters (clusters) containing possible resistance genes from whole genome sequencing

data. In order to do this we based the pipeline on the assumption that the resistance gene is found within a cluster and is a copy – a paralog – of an essential gene. This pattern and the resulting resistance mechanism has previously been described in two different cases in fungi [13, 16].

We have used complete and draft quality whole genome sequences, mainly from the 300 *Aspergillus* sequencing project [5, 6]. The input for the pipeline was chosen to consists of three different types of data derived from the whole genome sequence data: 1) predicted genes/proteins and functionally annotated proteins, for the functional annotation we used InterPro [19, 20]. 2) Predictions of secondary metabolite gene clusters, for this purpose we used a re-implementation of SMURF [21] as described in Vesth et al. [6]. 3) Groups of homologous protein sequences. We used a pipeline designed for *Aspergillus* data creating homologous protein families based on single linkage of bidirectional BLASTp hits (as described in [6]). It is assumed that proteins with similar, although not necessarily identical, function will be clustered into the same protein family. For our purpose this is useful, as resistance and target genes will be grouped into one family.

Using the described input the pipeline consists of a number of filtration steps designed to identify the most likely candidate clusters containing potential resistance genes, Figure 2. Several steps in the pipeline are designed to deal with and/or minimize the effect of possible errors in assembly and annotation due either to inherent errors in sequencing technologies or errors caused by the assembly and sequence quality of draft genomes. Several options have been added to the pipeline to allow the user to set the filters to allow for more or less noise caused by errors. Besides mitigating errors the filtering steps are also implemented to deal with biological diversity and differences.

Here we present the results of the pipeline using an extensive test dataset of 51 *Penicillium* and *Aspergillus* species containing a total of 3,276 predicted clusters and 26,551 secondary metabolite genes. The goal of the pipeline is to identify clusters containing resistance genes, assuming that the resistance genes are copies of essential genes. In this context, an essential gene is defined as a gene that has homologs in all the species included in the analysis.

Homolog count and 'Strict' cluster selection

The first step in the pipeline is to couple the homologous protein families to the predicted cluster genes, Figure 2 step 1. Following, the number of homologs – homology count – for each secondary metabolite gene in each organism/genome is identified. In addition, the number of homologs found in predicted clusters are recorded.

Next, clusters with potential resistance genes are selected based on a specific pattern of homology counts, Figure 2 step 2. In this step the user can select various levels of stringency for the selection pattern.

The most strict and simple selection pattern (Figure 2 step 2 left) identifies clusters where only one of the genes have a homolog in the genome, it can have only one such homolog, and this homolog must not be part of another cluster. The gene with the homolog is the presumed resistance gene and the homolog outside the cluster is the presumed target gene. Using this selection criteria on our test dataset, 262 clusters are identified, divided into 141 potential resistance genes protein families,

Figure 3A (dark turquoise). This corresponds to 8% of the total clusters in the data set.

This 'strict' selection pattern is very restrictive: if any one of the other cluster genes have a homolog anywhere in the genome, the cluster will be filtered away. Many clusters contain tailoring enzymes with common functions such as P450 or methyltransferases, these can be thought of as "household" functions in secondary metabolism. Clusters with common tailoring functions will therefore frequently have several homologs resulting in high homology counts and the cluster will therefore fall out of the 'strict' selection pattern. Another effect of the filtering is that the selected clusters often contain fewer genes (average size 6.4 genes vs. 8.2 for the total data set). The bigger the cluster, the more likely it is that there are several genes with homologs in the genome. It is likely that clusters containing potential resistance genes are missed due to the strict selection criteria. Conversely, the share of bioactive clusters is increased after the selection.

'Alternative' cluster selection pattern

To increase the number of selected clusters with potential resistance genes, we wanted to create an alternative selection pattern allowing the presence of more tailoring enzymes and larger cluster sizes. In order to generate reasonable alternative selection criteria, the cluster genes in the dataset were investigated. The most common InterPro domains (annotated in at least 1000 cluster genes) were selected, the protein families belonging to each InterPro domain was identified and the size of the protein families were noted. In the additional file C1 a violin plot illustrates the selected InterPro domains on the x-axis and the sizes of the protein families with at least one protein with this domain are on the y-axis. There are many protein families belonging to each InterPro domain ranging in size from 1 to 475 proteins.

Protein families with more than 100 members will most likely have several homologs in each organism and every time they appear in a cluster, the cluster will be discarded using the 'strict' selection pattern. To avoid this, an 'alternative' selection pattern was created disregarding large protein families as potential resistance genes and instead allowing the proteins belonging to large protein families to have homologs in the organism, Figure 2 step 2 'alternative' pattern.

The size of the protein families will change depending on the data (e.g. the number of genomes included and how closely related the species are). The selection was therefore designed to be dependent on the number of organisms in the data set. In order to determine this, we defined a metric: If the protein family is larger than the total number of organisms in the dataset multiplied by X, where X is a user input, then the gene is not considered as a potential resistance gene. Instead it is allowed to have homologs and the cluster can still be selected, if another gene meets the requirements for a resistance gene and the rest of the genes either have no homologs or belong to a large protein family. The alternative pattern under step 2 in Figure 2 illustrates the copy number pattern.

In the less strict selection method, the potential resistance gene is thus still only allowed to have one homolog outside the cluster and there is only one gene that can have this pattern, however the other genes in the cluster are allowed to have more homologs if their protein family is larger than: $X_{Input} \times Number\ of\ organisms$.

If X is set to 2 then the cutoff in the illustration (step 2 in Figure 2) would be 2*5 organisms and the protein family illustrated has 11 members and hence would be allowed in the pattern. If 3 is selected instead the cutoff would be 12 and the illustrated protein family is not big enough and this cluster would fall out.

We have used two and three as input thereby disregarding protein families with more than 102 or 153 members respectively. With this selection criteria 482 and 388 clusters are identified respectively which corresponds to an 84% and 48% increase compared to the initial 'strict' selection criteria, Figure 3A. Of these clusters there are 255 and 202 different potential resistance gene families, corresponding to an increase of 81% and 43% respectively, showing that the 'strict' measurement is indeed sensitive to large protein families.

Filtering

As mentioned above, genome data contains multiple types of errors such as incomplete genomes, gene calling and secondary metabolite gene cluster prediction errors. Incomplete genomes and gene calling errors are likely to casue false negatives in our pipeline while cluster prediction errors can cause both false positives and negatives depending on if the prediction algorithm over or under predicts the size of the cluster. Even with the current rate of technological advancement, data errors will most likely be a problem for some time to come and therefore we have created filtering steps to deal with these shortcomings in the data. Another reason for filtering is to identify the most likely resistance genes and hence filtering steps have also been implemented to decrease false positives.

Step 3 – Filtering the number of clusters. A filtering step is implemented in order to mitigate secondary metabolite prediction errors (Figure 2 step 3). With the assumption that if two of the selected clusters have potential resistance genes belonging to the same protein family, then it is less likely that it is an error and more likely that the resistance gene belongs to the cluster. The filtering step selects only the clusters where at least one other identified cluster has a potential resistance gene belonging to the same protein family. As such, the members of the potential resistance gene protein family have to be found in at least two selected clusters.

Using these filtering criteria, 45, 88, and 67 of the previous identified potential resistance genes protein families are left for the 'strict', 'alternative' X_{input} 2 and 3 which correspond to 32-35% of the initial selected potential resistance gene cases, Figure 3B. As such about a third of the initially identified clusters share resistance gene with another identified cluster and these are therefore more likely not to be prediction errors.

Step 4 – Filtering the number of organisms with homologs. The assumption behind this step is that the resistance genes should preferentially belong to a protein family with an essential function, and thus it should have protein members in all the species. The filtering step has two purposes, first it makes it more likely that it is a resistance gene if there are homologs in all species and second it will be a more widely useful bioactive compound if the target is conserved in many species.

This filtering step removes clusters where the putative resistance gene have homologs in less than a certain percentage of the organisms, Figure 2 step 4. The user

can select percentages from 100 to 90 to select the most likely resistance genes and best targets. The possibility of choosing lower than 100% was implemented to allow for some data and prediction errors such as incomplete genomes and imperfect gene annotations. Another reason for choosing lower than 100% is to allow some species to have a different mechanism and thus not the target gene. Here we show the effect of selecting 98% and 90% of the organism on the number of cases selected. If selecting the most restrictive setting where the homologs of the resistance gene has to be found in 98% of the organism, then there are 14, 25 and 22 potential resistance gene cases of the 'strict', 'alternative' X_{input} 2 and 3 respectively, Figure 3C. While setting the percentage of organisms to 90% there are 19, 38 and 32 potential resistance gene cases, Figure 3C.

If one chooses not to use step 3 (based on at least two selected clusters having the potential resistance gene), but only employ step 4 (based on the number of organisms homologs of the potential resistance gene should be found in) more clusters are selected, see Figure 3D. Using the setting of potential resistance genes having homologs in 98% of the organisms 43, 57 and 54 cases are identified based on the homology count pattern of 'strict', 'alternative' X_{input} 2 and 3, respectively. While employing only 90% of the organisms should have homologs of the potential resistance genes, 70, 99 and 89 are cases were detected for 'strict', 'alternative' X_{input} 2 and 3.

Step 5 – Filtering the number of organisms with single copies. The final filtering step is again related to the organisms in which the putative resistance gene have homologs. The assumption here is that the target gene should be one essential gene and therefore the majority of the species should only have one copy.

This filtering step therefore removes protein families where more than 50% of the organisms have more than one homolog of the putative resistance gene, Figure 2 step 5. This step is optional and can be employed or not as the user sees fit. The effect of adding this selection criteria can bee seen in Figure 3E and F.

Pipeline output The primary output of this pipeline is a list of protein families with potential resistance genes and fasta files containing all the proteins belonging to the identified family. The header of each entry includes the name of the organism, the section it belongs to, the protein id, the number of copies found in the organism and if it is in a selected cluster (StrictClust), a predicted cluster (Clust) or somewhere else in the genome (0) which is followed by the amino acid sequence. The cases that come out of the pipeline depends on the different settings and each identified case is independent of the others. If a high experimental capacity is available one option is to test all these cases, if the experimental capacity is limited further analysis is needed to evaluate and select which cases to work with.

In this example with 51 species, we get 12 different outputs after step 5 using each of the various settings. Narrowing down from 3,276 clusters we have identified 72 unique putative resistance gene families using all the various settings (Table C2). Of these 4 potential resistance gene families are found every time independent on which setting, while 19 are found only once using a specific set of variables.

With each filtering step, the number of clusters and putative resistance genes decreases, but the share of likely bioactive clusters and true resistance genes increases.

Identifying potential bioactive clusters

Even with the described selection criteria, the pipeline still produces more candidate clusters than it would be feasible for most academic labs to verify experimentally. Once the pipeline has been run and potential resistance gene cases have been identified, further analysis is needed to select the most promising candidates for experimental verification.

In order to do this efficiently, we use a combination of principal component analysis (PCA), phylogenetic analysis, functional annotation, and comparison to the NCBI database using BLASTp in order to gain more knowledge about the potential cases (examples of the PCA and phylogenetic trees can be seen in Figure 4 and additional Figure C2 and C2).

The functional annotation and BLAST analysis are included to add to the understanding of the potential resistance genes; to examine if it has a known function, if the function is essential, and if it would be a good drug target.

The PCA and phylogenetic trees are used to illustrate the evolutionary relationship between the proteins in the protein family. Resistance genes are expected to be a duplication of a target gene which is essential. Essential genes are under similar evolutionary pressure in all the species and are therefore closely related. The resistance gene however, is under a different pressure, potentially expressed under a different subset of conditions, which is expected to be reflected in the sequence and hence the analysis where resistance genes should form a separate group compared to the target genes. To perform these two analysis the protein sequences from the protein families were aligned using clustalo [22] and trimmed using Gblocks [23, 24].

As mentioned in the introduction, there are some known clusters containing a verified resistance gene, this includes the mycophenolic acid cluster in *P. brevicompactum* and the fellutamide B cluster from *A. nidulans*. Both clusters were included in this study as validation controls to check if the known resistance genes would be identified using our pipeline.

The Fellutamide B cluster is identified in all the outputs where step 3 was skipped, since this was the only cluster with this resistance gene and the selected homology count pattern. Thus when filtering for more reliable cluster predictions it falls out of the analysis.

The potential resistance and target genes are exposed to different evolutionary pressure. This is expected to be reflected in the evolutionary pattern where the essential target genes (under high pressure) will be closely related whereas the resistance genes will show more variation. To investigate if this is the case a phylogenetic tree and a principal component (PCA) analysis was performed for the protein family containing the fellutamide B resistance gene (596635), Figure 4. The PCA shows a clear distinction between the resistance genes found in clusters and the other essential target genes, Figure 4B. The phylogenetic tree (Figure 4A) shows two clear groupings: one big closely related group with all the target genes (having 0 at the end of the label), and one small containing the resistance gene. The target genes are ordered in their phylogenetic groups, so species from the same section cluster together. The resistance gene from *A. nidulans* is clustered with a protein from *A. sydowii* and *A. versicolor*, which both are found in clusters, indicating

that these most likely also function as resistance genes and potentially can produce Fellutamide B or a derivative or a different compound attacking the same target. *A. versicolor* is known to produce fellutamide C and F [25, 26], whereas *A. sydowii* to our knowledge has not been reported to produce derivatives of fellutamide. The clusters in *A. versicolor* and *A. sydowii* are similar to the *A. nidulans* fellutamide B cluster with similar backbone enzymes ($\geq$ 70%) and tailoring enzymes, but the *A. versicolor* and *A. sydowii* predicted clusters are bigger consisting of 19 genes, 7 more than the *A. nidulans* cluster. Three of the extra genes have homologs in the genome and belong to small protein families which is why these clusters are not identified in the first steps of the pipeline.

Near the resistance genes in the phylogenetic tree, there are also three other genes found in *A. homomorphus*, *A. tachungensis* and *A. candidus* which are extra copies of the target genes, but they are not predicted to be in clusters. One explanation for the extra copies not found in clusters could be cluster prediction errors. We therefore investigated if *A. homomorphus*, *A. tachungensis* and *A. candidus* had a cluster similar to the *A. nidulans* fellutamide B cluster, but no BLASTp hits of the backbone genes had identity above 37% indicating that they most likely cannot produce any derivatives of fellutamide. Another explanation for the extra copy could be that it functions as a defence mechanism protecting against other species producing fellutamide, which could be useful if the species naturally grow near fellutamide producing species.

The cluster of mycophenolic acid did not turn up in our outputs, so we investigated the cluster further to understand why. The predicted cluster containing the PKS responsible for producing the core compound of mycophenolic acid consists of four genes; The PKS, a P450, a methyltransferase and the resistance gene. The PKS and p450 are only found in one copy and the resistance gene has one identified homolog as expected. The methyltransferase however also has a homolog in the genome. As the size of this protein family is only 23 members, it is not disregarded in the alternative pattern and the cluster is therefore filtered away in step 2. This shows that we do loose some good cases along the steps in the pipeline which illustrates the importance of running the pipeline with multiple settings and inspecting the output carefully. The pipeline is highly sensitive to the number of organisms included and the settings should be used keeping this limitation in mind.

Novel putative resistance gene

In addition to the known clusters with resistance genes, several uncharacterized clusters were also identified containing putative resistance genes. We will here focus on one example where the potential resistance gene is found in clusters in *A. oryzae* and *A. flavus* (protein family 597268). The PCA analysis showed a very clear picture of the resistance genes falling outside the group of target genes, the same is seen in the phylogenetic tree, where the target genes also follow the expected phylogeny with species from the same section grouping together, Additional Figure C2 and C3. Besides the identified putative resistance genes in *A. oryzae* and *A. flavus* several other species (*A. wentii, A. piperis, A. candidus, A. taichungensis, A. campestris, A. novofumigatus* and *P. brevicompactum*) have an additional gene but these are not found in predicted clusters. The backbone genes from *A. oryzae* and *A. flavus*

have no hits in those species which could mean that the species carries the resistance gene but do not produce the compound.

The predicted clusters in *A. oryzae* and *A. flavus* both consists of four genes: an acetyltransferase (IPR000182, IPR016181), the predicted reistance gene, a NRPS-like synthetase and a gene belonging to the major facilitator superfamily (IPR011701, IPR007114). The putative resistance gene has annotations involved in 'Signal transduction response regulator, receiver region' and 'Signal transduction histidine kinase, core' (IPR001789, IPR005467). Using BLASTp to investigate the function of the putative resistance gene, the top hits have functions like: 'two-component osmosensing histidine kinase (Bos1)' (RAQ55620.1). Based on this information it seem likely we have identified a cluster containing a previously unidentified resistance gene where the compound inhibits a histidine kinase. In order to fully determine this, this will have to be experimentally validated in the future.

Discussion

The pipeline was designed to identify secondary metabolite gene clusters (clusters) containing potential resistance genes. It is a delicate balance of filtering away as many clusters as possible to narrow down the field to the best candidates, while keeping as many clusters with potential resistance gene as possible.

We have chosen an approach, where we make no assumptions about the function or makeup of resistance genes besides being homologs of a gene shared by most organisms in the data set. Another approach could be to screen for specific classes of essential genes or resistance targets within predicted clusters. This approach was used in another study where 86 bacterial *Salinispora* genomes were mined for duplicated genes involved in central metabolism co-localizing with clusters. Clusters containing putative fatty acid synthase resistance genes were identified and these were shown to be involved in the biosynthesis of thiotetronic acid natural products, including thiolactomycin which is a well-known fatty acid synthase inhibitor [27]. This approach builds on knowledge of house-keeping genes thus requiring extensive knowledge of the primary metabolism. In filamentous fungi there is still a lot of primary processes which are not characterized, therefore there is a risk of missing interesting resistance genes using that approach. To avoid this, we chose to use a wider approach based only on the homology copy number pattern of predicted secondary metabolism genes and not based on the functions. The underlying assumption is that our pipeline identifies conserved household genes with a homolog in a cluster. Using our approach, we avoid limiting the search space to only known mechanisms thereby making it possible to find new essential mechanisms and drug targets. Our non-functional-impelled approach can be used both on organisms with little knowledge and extensive knowledge of the primary metabolism. In well characterized species our approach is also useful since it is likely both to identify known household genes but also new uncharacterized household genes. Finally, the setup makes it possible to search the identified clusters afterwards with a criterion such as the presence of primary metabolism genes.

The pipeline has been designed for a specific data set-up but it is also possible to apply the method and the approach to other data-sets using the same ratiocination.

In each of the steps in the pipeline various parameters affect the output and identified putative resistance genes, which parameters to use and tweak depends

on the data and the aim of the analysis. The first thing is therefore to select the input data carefully. Data with distantly related species might not give meaningful homologous protein families using our cut-offs since proteins with similar function might be more different than our cut-offs allow. Step 1 in the pipeline combines the input data and creates a tables with homology count of all the cluster genes, in this step no filtering or selection is done and hence there is no parameters to tweak.

In step 2 selecting clusters with a specific homology count pattern, the strict or alternative patterns can be selected. Using the strict selection criteria, only 8% of the clusters meets the criteria. When using too restrictive selection patterns, there is a risk of creating false negatives, thus filtering away good cases. As mentioned earlier, many common secondary metabolite genes are likely to have homologs and belong to large protein families. Of the cluster genes belonging to this test dataset 12 and 6% belong to protein families larger than 102 and 153 proteins respectively, and these are found in 51% and 37% of the total clusters. Therefore using the 'strict' selection pattern, up to half of the clusters are likely to be discarded due to the homology count pattern of standard cluster genes, which causes a lot of false negatives. We therefore recommend using the alternative selection pattern. If the species in the dataset are distantly related the size of the protein families might be smaller and therefore a lower $Xinput$ is recommended. The lowest reasonable value of $Xinput$ we recommend is 1.5; going lower the risk of disregarding true resistance gene families becomes too big.

Filtering based on more clusters having the putative resistance gene (step 3) was designed to avoid false positives due to cluster prediction errors. This step is useful for data including closely related species likely to have similar clusters. If the data consists mainly of distantly related species, it is less likely that similar clusters will be found in more species and hence the risk of filtering away good cases increases thereby making false negatives. In the test data, about two-thirds of the cases are filtered away in this step, if running the pipeline and even more than two-tirds are filtered away, we suggest skipping this step.

The 4th filtering step was made to avoid false positives, in this step the percentage of organisms that should have a homolog of the putative resistance gene is selected and this parameter can be tweaked depending on the data. If the species are distantly related, it is more likely that some species have a different essential household mechanism and hence does not have a copy the household / target gene. A lower percentage of organisms that should have a homolog is thus recommended in this case. Another reason for choosing a lower percentage of organisms with a homolog is if the quality of the genome sequence data is low, with incomplete genomes or if the data includes novel species distantly related to model organisms where the gene prediction algorithms might not work as efficiently.

The last filtering step was also made to avoid false positives, in this step at least 50% of the species should only have a single homolog. This filtering step is independent of the quality of the data and the relatedness of the species, and we therefore recommend using this at all times. Here 30-60% of the cases are left after this filtering, significantly decreasing the number but leaving highly promising resistance cases.

Conclusions

In this study, we have created a method identifying clusters responsible for producing bioactive compounds. The approach we have developed is based on a specific resistance mechanism and paves the way for rationally selecting promising bioactive clusters from whole genome data. The FRIGG pipeline was designed in connection with the *Aspergillus* sequencing project however several filtering steps and paramters can be tweaked to fit different kinds of data and to deal with the most likely errors from predictions and draft genomes.

We have tested the developed pipeline on 51 *Aspergillus* and *Penicillium* genomes identifying 72 unique putative resistance genes and clusters in the most strict configuration of the pipeline. In addition, the characterized Fellutamide B resistance gene *inpE* was successfully identified with this pipeline confirming the accuracy and applicability of the pipeline to such cases of resistance mechanisms.

As more and more genomes are sequenced, the relevance of this approach will increase and it will become a useful method for selecting which clusters to focus on in the hunt for novel drugs such as anti-fungicides, anti-cancer drugs and antimicrobial compounds.

Methods

Fungal species

The data consisted of 50 *Aspergillus* and 1 *Penicillium* species with available whole genome sequencing data which was downloaded from JGI. Species information can be found in Additional Table C1.

Input data

The data used in the pipeline was organized in a MySQL database, an overview of the input data can be found in the file: Input_data_pipeline.txt and the data can be found as sql/csv files on https://github.com/ingek-1/FRIGG_pipeline. A few of the data files were too big for Github and can be found in this repository instead https://files.dtu.dk/u/ox6SPjaekiyxHpI8/Data_tables_FRIGG?l. The data includes predicted secondary metabolite gene clusters based on an implementation of the SMURF [21] pipeline described elsewhere [6]. Homologous protein families were created with Aspergillus optimized parameters based on single linkage of bidirectional BLASTp hits with identity $\geq 50\%$ and sum of query and hit coverage $\geq 130\%$ as described in [6]. InterPro annotations of the proteins were also used [19, 20] and gff information.

Pipeline

The pipeline was created using Python and the investigation of specific protein families was conducted in R [28]. The versions and packages used can be seen in version info files on the github page. For alignment and trimming of the protein families clustalo and Gblocks was used in a Python script, also included on the github page. All the scripts and files are available at Github: https://github.com/ingek-1/FRIGG_pipeline.

Competing interests
The authors declare that they have no competing interests.

Author's contributions

Acknowledgements

References

1. Keller, N.P., Turner, G., Bennett, J.W.: Fungal secondary metabolism – from biochemistry to genomics. Nature Reviews Microbiology **3**(12), 937–947 (2005). doi:10.1038/nrmicro1286
2. Galagan, J.E., Calvo, S.E., Cuomo, C., Ma, L.-J., Wortman, J.R., Batzoglou, S., Lee, S.-I., Baştürkmen, M., Spevak, C.C., Clutterbuck, J., Kapitonov, V., Jurka, J., Scazzocchio, C., Farman, M., Butler, J., Purcell, S., Harris, S., Braus, G.H., Draht, O., Busch, S., D'Enfert, C., Bouchier, C., Goldman, G.H., Bell-Pedersen, D., Griffiths-Jones, S., Doonan, J.H., Yu, J., Vienken, K., Pain, A., Freitag, M., Selker, E.U., Archer, D.B., Peñalva, M.A., Oakley, B.R., Momany, M., Tanaka, T., Kumagai, T., Asai, K., Machida, M., Nierman, W.C., Denning, D.W., Caddick, M., Hynes, M., Paoletti, M., Fischer, R., Miller, B., Dyer, P., Sachs, M.S., Osmani, S.A., Birren, B.W.: Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae. Nature **438**(7071), 1105–15 (2005). doi:10.1038/nature04341
3. Machida, M., Asai, K., Sano, M., Tanaka, T., Kumagai, T., Terai, G., Kusumoto, K.-I., Arima, T., Akita, O., Kashiwagi, Y., Abe, K., Gomi, K., Horiuchi, H., Kitamoto, K., Kobayashi, T., Takeuchi, M., Denning, D.W., Galagan, J.E., Nierman, W.C., Yu, J., Archer, D.B., Bennett, J.W., Bhatnagar, D., Cleveland, T.E., Fedorova, N.D., Gotoh, O., Horikawa, H., Hosoyama, A., Ichinomiya, M., Igarashi, R., Iwashita, K., Juvvadi, P.R., Kato, M., Kato, Y., Kin, T., Kokubun, A., Maeda, H., Maeyama, N., Maruyama, J.-I., Nagasaki, H., Nakajima, T., Oda, K., Okada, K., Paulsen, I., Sakamoto, K., Sawano, T., Takahashi, M., Takase, K., Terabayashi, Y., Wortman, J.R., Yamada, O., Yamagata, Y., Anazawa, H., Hata, Y., Koide, Y., Komori, T., Koyama, Y., Minetoki, T., Suharnan, S., Tanaka, A., Isono, K., Kuhara, S., Ogasawara, N., Kikuchi, H., The: Genome sequencing and analysis of Aspergillus oryzae. Nature **438**, 1157–1161 (2005). doi:10.1038/nature04300
4. Nierman, W.C., Pain, A., Anderson, M.J., Wortman, J.R., Kim, H.S., Arroyo, J., Berriman, M., Abe, K., Archer, D.B., Bermejo, C., Bennett, J., Bowyer, P., Chen, D., Collins, M., Coulsen, R., Davies, R., Dyer, P.S., Farman, M., Fedorova, N., Fedorova, N., Feldblyum, T.V., Fischer, R., Fosker, N., Fraser, A., García, J.L., García, M.J., Goble, A., Goldman, G.H., Gomi, K., Griffith-Jones, S., Gwilliam, R., Haas, B., Haas, H., Harris, D., Horiuchi, H., Huang, J., Humphray, S., Jiménez, J., Keller, N., Khouri, H., Kitamoto, K., Kobayashi, T., Konzack, S., Kulkarni, R., Kumagai, T., Lafton, A., Latgé, J.-P., Li, W., Lord, A., Lu, C., Majoros, W.H., May, G.S., Miller, B.L., Mohamoud, Y., Molina, M., Monod, M., Mouyna, I., Mulligan, S., Murphy, L., O 'neil, S., Paulsen, I., Peñalva, M.A., Pertea, M., Price, C., Pritchard, B.L., Quail, M.A., Rabbinowitsch, E., Rawlins, N., Rajandream, M.-A., Reichard, U., Renauld, H., Robson, G.D., Rodriguez De Córdoba, S., Rodríguez-Peña, J.M., Ronning, C.M., Rutter, S., Salzberg, S.L., Sanchez, M., Sánchez-Ferrero, J.C., Saunders, D., Seeger, K., Squares, R., Squares, S., Takeuchi, M., Tekaia, F., Turner, G., Vazquez De Aldana, C.R., Weidman, J., White, O., Woodward, J., Yu, J.-H., Fraser, C., Galagan, J.E., Asai, K., Machida, M., Hall, N., Barrell, B., Denning, D.W.: Genomic sequence of the pathogenic and allergenic filamentous fungus Aspergillus fumigatus. Nature **438**, 1151–1156 (2005). doi:10.1038/nature04332
5. Kjærbølling, I., Vesth, T.C., Frisvad, J.C., Nybo, J.L., Theobald, S., Kuo, A., Bowyer, P., Matsuda, Y., Mondo, S., Lyhne, E.K., Kogle, M.E., Clum, A., Lipzen, A., Salamov, A., Ngan, C.Y., Daum, C., Chiniquy, J., Barry, K., LaButti, K., Haridas, S., Simmons, B.A., Magnuson, J.K., Mortensen, U.H., Larsen, T.O., Grigoriev, I.V., Baker, S.E., Andersen, M.R.: Linking secondary metabolites to gene clusters through genome sequencing of six diverse Aspergillus species. Proceedings of the National Academy of Sciences **115**(4), 753–761 (2018). doi:10.1073/pnas.1715954115
6. Vesth, T.C., Nybo, J.L., Theobald, S., Frisvad, J.C., Larsen, T.O., Nielsen, K.F., Hoof, J.B., Brandl, J., Salamov, A., Riley, R., Gladden, J.M., Phatale, P., Nielsen, M.T., Lyhne, E.K., Kogle, M.E., Strasser, K., McDonnell, E., Barry, K., Clum, A., Chen, C., LaButti, K., Haridas, S., Nolan, M., Sandor, L., Kuo, A., Lipzen, A., Hainaut, M., Drula, E., Tsang, A., Magnuson, J.K., Henrissat, B., Wiebenga, A., Simmons, B.A., Mäkelä, M.R., de Vries, R.P., Grigoriev, I.V., Mortensen, U.H., Baker, S.E., Andersen, M.R.: Investigation of inter- and intra-species variation through genome sequencing of Aspergillus section Nigri. Nature Genetics **Accepted** - (2018)
7. Demain, A.L.: How Do Antibiotic-Producing Microorganisms Avoid Suicide? Annals of the New York Academy of Sciences **235**(1), 601–612 (1974). doi:10.1111/j.1749-6632.1974.tb43294.x
8. Cundliffe, E.: HOW ANTIBIOTIC-PRODUCING ORG\.NISMS A VOID SUICIDE. Technical report (1989). www.annualreviews.org
9. Hopwood, D.A.: How do antibiotic-producing bacteria ensure their self-resistance before antibiotic biosynthesis incapacitates them? Molecular Microbiology **63**(4), 937–940 (2007). doi:10.1111/j.1365-2958.2006.05584.x. NIHMS150003
10. Steffensky, M., Mühlenweg, A., Wang, Z.-x., Li, S.-m., Mu, A., Heide, L.: Identification of the Novobiocin Biosynthetic Gene Cluster of Streptomyces spheroides NCIB 11891. Antimicrobial Agents and Chemotherapy **44**(5), 1214–1222 (2000). doi:10.1128/AAC.44.5.1214-1222.2000.Updated
11. Frohlich, K.U., Wiemann, M., Lottspeich, F., Mecke, D.: Substitution of a pentalenolactone-sensitive glyceraldehyde-3-phosphate dehydrogenase by a genetically distinct resistant isoform accompanies pentalenolactone production in Streptomyces arenae. Journal of Bacteriology **171**(12), 6696–6702 (1989). doi:10.1128/jb.171.12.6696-6702.1989
12. Tetzlaff, C.N., You, Z., Cane, D.E., Takamatsu, S., Omura, S., Ikeda, H.: A gene cluster for biosynthesis of the sesquiterpenoid antibiotic pentalenolactone in Streptomyces avermitilis. Biochemistry **45**(19), 6179–6186 (2006). doi:10.1021/bi060419n. NIHMS150003
13. Hansen, B.G., Genee, H.J., Kaas, C.S., Nielsen, J.B., Regueira, T.B., Mortensen, U.H., Frisvad, J.C., Patil, K.R.: A new class of IMP dehydrogenase with a role in self-resistance of mycophenolic acid producing fungi. BMC microbiology **11**, 202 (2011). doi:10.1186/1471-2180-11-202
14. Sun, X.E., Hansen, B.G., Hedstrom, L.: Kinetically Controlled Drug Resistance. Journal of Biological Chemistry **286**(47), 40595–40600 (2011). doi:10.1074/jbc.m111.305235

15. Hansen, B., Sun, X., Genee, H., Kaas, C., Nielsen, J., Mortensen, U., Frisvad, J., Hedstrom, L.: Adaptive evolution of drug targets in producer and non-producer organisms. Biochemical Journal **441**(1), 219–226 (2012). doi:10.1042/BJ20111278
16. Yeh, H.-H., Ahuja, M., Chiang, Y.-M., Oakley, E., Moore, S., Yoon, O., Hajovsky, H., Bok, J.W., Keller, N.P., Wang, C.C.C., Oakley, B.R.: Resistance gene-guided genome mining: Serial promoter exchanges in Aspergillus nidulans reveal the biosynthetic pathway for fellutamide B, a proteasome inhibitor. ACS Chemical Biology **11**(8), 2275–2284 (2016). doi:10.1021/acschembio.6b00213
17. Cochrane, R.V.K., Sanichar, R., Lambkin, G.R., Reiz, B., Xu, W., Tang, Y., Vederas, J.C.: Production of New Cladosporin Analogues by Reconstitution of the Polyketide Synthases Responsible for the Biosynthesis of this Antimalarial Agent. Angewandte Chemie - International Edition **55**(2), 664–668 (2016). doi:10.1002/anie.201509345
18. Yan, Y., Liu, Q., Zang, X., Yuan, S., Bat-Erdene, U., Nguyen, C., Gan, J., Zhou, J., Jacobsen, S.E., Tang, Y.: Resistance-gene-directed discovery of a natural-product herbicide with a new mode of action. Nature **559**(7714), 415–418 (2018). doi:10.1038/s41586-018-0319-4
19. Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A.F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.Y., Lopez, R., Hunter, S.: InterProScan 5: Genome-scale protein function classification. Bioinformatics **30**(9), 1236–1240 (2014). doi:10.1093/bioinformatics/btu031
20. Finn, R.D., Attwood, T.K., Babbitt, P.C., Bateman, A., Bork, P., Bridge, A.J., Chang, H.Y., Dosztanyi, Z., El-Gebali, S., Fraser, M., Gough, J., Haft, D., Holliday, G.L., Huang, H., Huang, X., Letunic, I., Lopez, R., Lu, S., Marchler-Bauer, A., Mi, H., Mistry, J., Natale, D.A., Necci, M., Nuka, G., Orengo, C.A., Park, Y., Pesseat, S., Piovesan, D., Potter, S.C., Rawlings, N.D., Redaschi, N., Richardson, L., Rivoire, C., Sangrador-Vegas, A., Sigrist, C., Sillitoe, I., Smithers, B., Squizzato, S., Sutton, G., Thanki, N., Thomas, P.D., Tosatto, S.C.E., Wu, C.H., Xenarios, I., Yeh, L.S., Young, S.Y., Mitchell, A.L.: InterPro in 2017-beyond protein family and domain annotations. Nucleic Acids Research **45**(D1), 190–199 (2017). doi:10.1093/nar/gkw1107
21. Khaldi, N., Seifuddin, F.T., Turner, G., Haft, D., Nierman, W.C., Wolfe, K.H., Fedorova, N.D.: SMURF: Genomic mapping of fungal secondary metabolite clusters. Fungal Genetics and Biology **47**(9), 736–741 (2010). doi:10.1016/j.fgb.2010.06.003
22. Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., Mcwilliam, H., Remmert, M., Sö Ding, J., Thompson, J.D., Higgins, D.G.: Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology **7**, 539 (2011). doi:10.1038/msb.2011.75
23. Castresana, J.: Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. Molecular Biology and Evolution **17**(4), 540–552 (2000)
24. Talavera, G., Castresana, J.: Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments. Systematic Biology **56**(4), 564–577 (2007). doi:10.1080/10635150701472164
25. Lee, Y.M., Dang, H.T., Li, J., Zhang, P., Hong, J., Lee, C.O., Jung, J.H.: A cytotoxic fellutamide analogue from the sponge-derived fungus aspergillus versicolor. Bulletin of the Korean Chemical Society **32**(10), 3817–3820 (2011). doi:10.5012/bkcs.2011.32.10.3817
26. Zhang, H., Zhao, Z., Wang, H.: Cytotoxic natural products from marine sponge-derived microorganisms. Marine Drugs **15**(3) (2017). doi:10.3390/md15030068
27. Tang, X., Li, J., Millán-Aguiñaga, N., Zhang, J.J., O'Neill, E.C., Ugalde, J.A., Jensen, P.R., Mantovani, S.M., Moore, B.S.: Identification of Thiotetronic Acid Antibiotic Biosynthetic Pathways by Target-directed Genome Mining. ACS Chemical Biology **10**(12), 2841–2849 (2015). doi:10.1021/acschembio.5b00658
28. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2018). R Foundation for Statistical Computing. https://www.R-project.org/

Figures

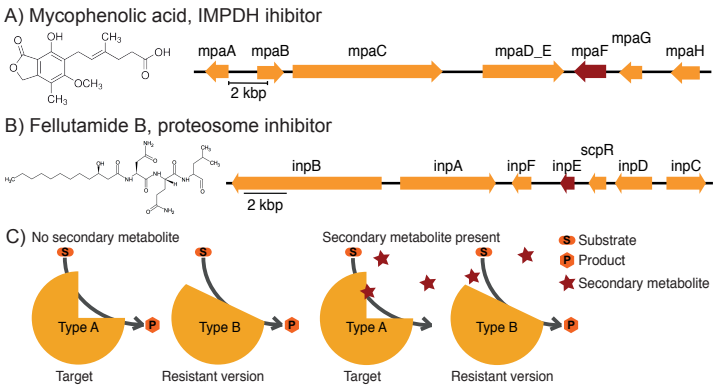


Figure 1 Resistance mechanism. A) Mycophenolic acid chemical structure and biosynthetic cluster with known resistance gene mpaF (highligted in red) which is a Inosine-5'-monophosphate dehydrogenase (IMPDH) inhibitor. B) Chemical struture of Fellutamide B and overview of the biosynthetic cluster including the resistance gene inpE (highlighted in red) which is a proteasome inhibitor. C) Illustration of the resistance mechanism used by some toxin producers. The secondary metabolite is a toxin which inhibits an essential enzyme - the target of the compound. Within the cluster responsible for producing the toxin a copy of the target gene is found, this version is still functioning despite the compounds pressence and hence makes the organism self-resistant.

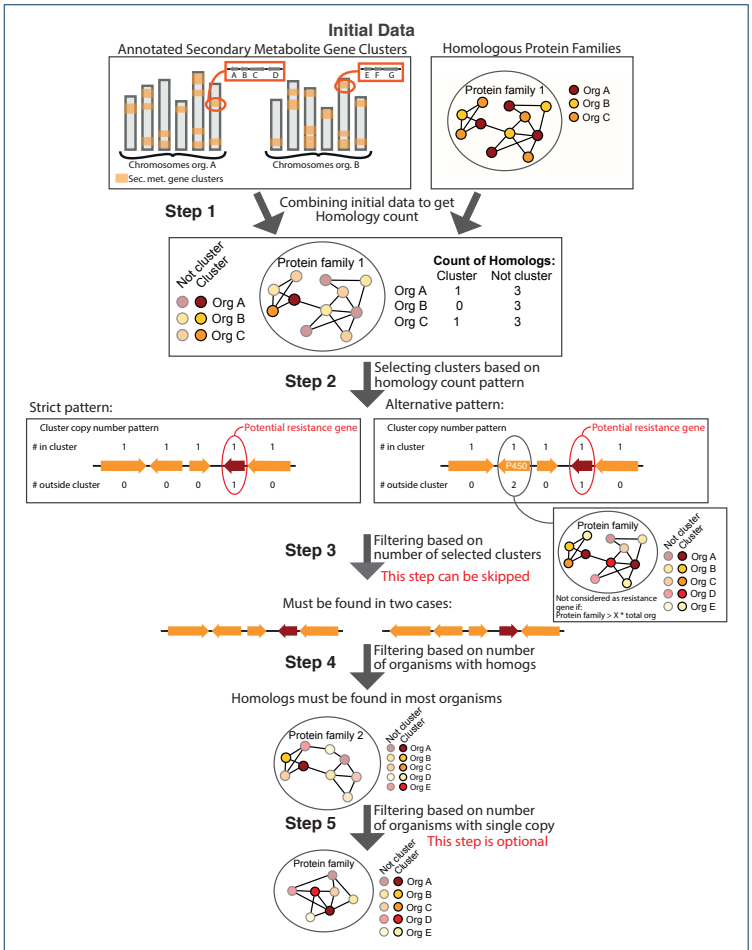


Figure 2 Overview of the pipeline. Overview of the pipeline illustrating the initial data of predicted secondary metabolite gene clusters and homologous protein families. Step 1 – The initial data is combined to generate counts of how many homologs are found in each organisms andhow many are within predicted clusters. Step 2 – Clusters are selected based on specific homology count patterns, either 'strict' where only one gene can have a homolog outside the cluster or 'Alternative' where genes belonging to large protein families are allowed. Step 3 – Filtering of the selected clusters based on other clusters having a homologous resistance gene, this step is optional. Step 4 – Filtering based on the majority of the organisms should have a homolog of the resistance/target genes. Step 5 – Filtering based on the number of organism that should only have a single homolog.

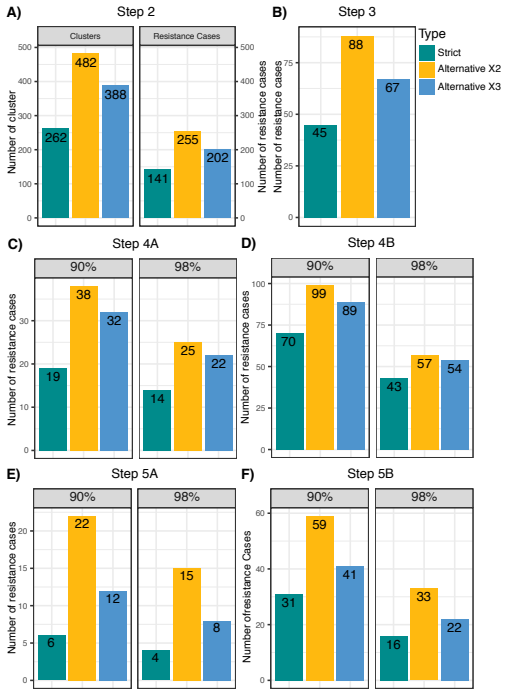


Figure 3 Number of cases after each filtering step using various settings. A) the number of cluster selected for each of the selections patterns and the number of resistance cases. B) The number of resistance cases after step 3 - the resistance protein family has to be found in two selected clusters. C) The number of resistance cases after step 4 - the resistance protein family have to have homologs in 90 or 98% of the organisms in the dataset. D) The number of resistance cases after step 4 but having skipped step 3, again both for 90 and 98% of the organisms. E) The number of resistance cases after step 5 - the resistance protein family have to be found as single copy in 50% of the species. F) The number of resistance cases after step 5 but having skipped step 3.

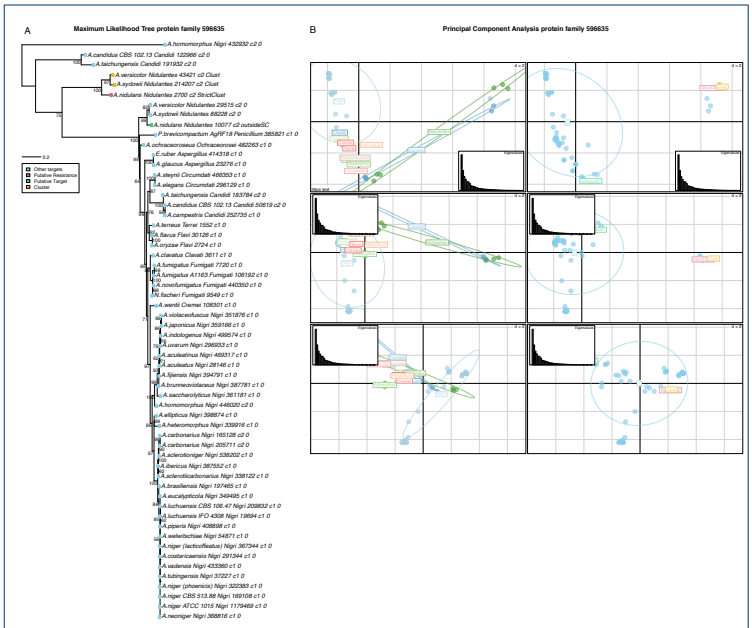


Figure 4 Phylogenetic tree and principal component analysis of protein family 596635 including the Fellutamide B resistance gene. A) Phylogenetic tree of the the protein family (596635) containing the Fellutamide B resistance gene, (protein id 2700) from *A. nidulans*. The phylogenetic tree is a maximum likelihood tree and was created with 500 bootstraps. The identified resistance gene is highlighted with red, the corresponding target is in green, other cluster genes are in yellow and gene not found in clusters are in blue. B) Principal component analysis of the protein family 596635 containing the Fellutamide B resistance gene, (protein id 2700) from *A. nidulans*. The panels to the left is colored based on the sections the proteins belong to while the panels to the right is colored based on if the protein if found in a a selected cluster (StrictClust- red), a predicted cluster (Clust-yellow), not in a cluster (0-blue) and the homolg to the ones found in selected clusters (outsideSC-green).

1 Additional Files

Figure C1 Common InterPro domains in secondary metabolism. Visualization of the most common InterPro annotations of secondary metabolite genes (found in more than 1000 secondary metabolite proteins) and the size of the protein families. Two horizontal lines indicate the recommended protein family size cut-offs where X_{Input} is 2 (102) and 3 (153).

Figure C2 Principal component analysis of the protein family 597268. Principal component analysis of the protein family 597268 containing two potential resistance genes (protein id 11595 and 32200) found in *A. oryzae* and *A. flavus*. The panels to the left are colored based on the sections the proteins belong to while the panels to the right are colored based on if the protein if found in a a selected cluster (StrictClust- green), not in a cluster (0-blue), and the homolg to the ones found in selected clusters (outsideSC-yellow).

Figure C3 Phylogenetic tree of the the protein family (597268). Phylogenetic tree of protein family 597268 containing two potential resistance genes (protein id 11595 and 32200) found in *A. oryzae* and *A. flavus*. The node labels shows bootstraps values based on 500. The tip labels have the species name, the section, protein id, number of homologs in the species and an indication if it is found in a selected cluster (StrictClust), a predicted cluster (Clust), not in a cluster (0) and the homolg to the ones found in selected clusters (outsideSC)

Table C1 Species used in this study, showing species name, section, and link to the JGI pages with the genomes.

Table C2 Overview of the 72 identified putative resistance genes families and the parameters where they were identified.

4.2 Identified and investigated resistance case

4.2.1 Introduction

Using the pipeline described in section 4.1, several clusters with putative resistance genes were identified. Out of these, one cluster was selected for further experimental analysis in order to verify that the putative resistance gene truly is a resistance gene and to identify the compound produced by the cluster. The additional files can be found in Appendix D.

Investigation of an *A. oryzae* predicted cluster

The selected putative resistance gene belongs to protein family (597268) also mentioned in section 4.1, Manuscript III. It is found in clusters in *A. flavus* and *A. oryzae* (no homolog of the cluster is found found in *A. nidulans* or *A. aculeatinus*). The predicted cluster consists of four genes in both species. The illustration in Figure 4.1 shows a representation of the *A. oryzae* cluster, but the cluster in *A. flavus* is syntenic. The predicted functions in the cluster: a major facilitator superfamily (IPR007114, IPR011701, IPR016196, JGI protein ID 11593), an NRPS-like protein (JGI protein ID 11594), a signal transduction histidine kinase /Signal transduction response regulator, receiver region (IPR001789/IPR005467, JGI protein ID 11595) which is the putative resistance gene, and an N-acetyltransferase/ Acyl-CoA N-acyltransferase (IPR000182, IPR016181, JGI protein ID 11596). It is expected that the NRPS-like and the N-acetyltransferase are involved in the biosynthesis of the secondary metabolite. The function of the major facilitator is not known, it could be responsible for transporting precursors or intermediates to the right compartments, or transporting the final compound out of the cell and potentially helping in self-protection as mentioned in section 2.3.

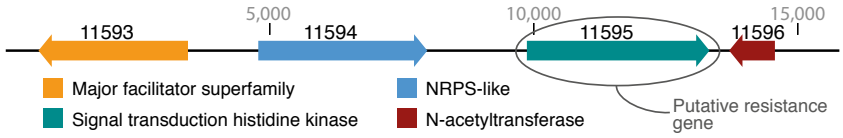


Figure 4.1 Selected cluster in *A.oryzae* with putative resistance gene. The genes are represented by arrows and the JGI protein id is shown above, while the functions are illustrated by the colors.

In order to investigate the cluster we have used several different strategies, I)

Gene deletion in the native host using classical gene targeting and CRISPR based deletion, II) Heterologous expression in *A. nidulans* by targeted integration, and III) random integration in *A. aculeatinus*.

As mentioned in the review in Section 2.4, the main advantage of using the native host is that clear changes can be observed in the chemical profile if the cluster is active. Classical gene targeting is dependent on the homologous recombination efficiency which in filamentous fungi can be low. More recently, CRISPR-Cas9 has been used to increase gene targeting efficiency by creating double stranded breaks in the locus of interest which then need to be repaired either by non-homologous end joining or homologous recombination using a plasmid or oligo as repair template [104, 105]. If the strain used is deficient in non-homologous end joining (NHEJ) the strain is forced to repair the breaks by homologous recombination (HR) thereby significantly increasing the gene targeting efficiency.

Heterologous expression in filamentous fungi can be used if it is not possible to use the native host, this strategy is laborious and have the challenges that the host can be affected by the inserted cluster affecting the chemical profile or the host can make cross-chemistry with the compound of interest making it difficult to identify the compound of interest.

4.2.2 Results and discussion

Strategy I: gene deletion in *A. oryzae*

Similar clusters were identified in both *A. flavus* and *A. oryzae*, but since we had more information and tools available for *A. oryzae* we chose to perform the experiments in that species.

From available transcriptomic data, it was evident that the cluster genes are expressed under standard laboratory conditions [106, 107]. We therefore decided to use a strategy of gene deletion of the cluster genes. We expected that we would be able to identify the compound produced by the cluster by comparing the chemical profile of the wild type (WT) and deletion mutants. In addition, we anticipated that we would be able to test if the putative resistance gene truly confers resistance to a yet unknown compound.

An available strain of *A. oryzae* engineered for molecular manipulation was used; *A. oryzae* A1560 ($\Delta pyrG$, $\Delta ku70$). This strain has the *pyrG* gene, coding for orotidine 5'-phosphate decarboxylase deleted, which is used as an auxotrophic marker, and it is deficient in non-homologous end-joining thus forcing the DNA repair to be done using homologous recombination (HR).

Vectors designed to delete cluster genes were constructed with *A. flavus pyrG* flanked by approximately 1000-base pairs targeting sequences complementary to up- and downstream regions of either the NRPS-like (plasmid pFC873), putative resistance alone (plasmid pFC874), or of both genes together (plasmid pFC963), overview of plasmids in appendix D Table D.3. Transformation of the *A. oryzae* A1560 ($\Delta pyrG$, Δku) with linearized vector pFC874, which should delete the putative resistance gene, resulted in extremely small colonies not viable and therefore unable to verify correct transformation. The transformants with linearized pFC873 and pFC963 were tested by analytical tissue PCR, but none had the correct deletions.

After several transformation attempts using classical gene targeting strategy, a method based on CRISPR-Cas9 was employed in order to increase the efficiency of homologous recombination (HR) [104, 105]. With HR, double-stranded breaks are induced at the target site forcing the fungi to repair the breaks either by NHEJ or HR if a repair template is given. NHEJ deficient strains must use a repair template and HR which should increase the efficiency of gene targeting.

The efficiency of the methods were tested by targeting the spore pigment color gene, *fwnA*, which when deleted gives white colonies instead of fawn. The efficiency of both standard gene targeting and CRISPR-based methods were tested in two strains *A. oryzae* A1560 $\Delta pyrG$, Δku and *A. oryzae* A1560 $\Delta pyrG$. This study showed that gene targeting in *A. oryzae* A1560 $\Delta pyrG$, Δku was unsuccessful, since no colonies grew, while the same strategy resulted in one big white colony in the *A. oryzae* A1560 $\Delta pyrG$ strain which is not what we expected (Figure 4.2). The strain with the *ku* deletion should be more efficient in HR than the strain with the wild-type *ku*. Co-transformation of a CRISPR-Cas9 plasmid carrying 2 gRNAs targeting the pigment gene in each end, with 90 bp oligo nucleotide repair fragment complementary to up- and downstream sequences of the cut sites, worked well in both strains giving several white colonies, Figure 4.2, column 3. Encouraged by these results confirming that the method is efficient in *A. oryzae*, new CRISPR-Cas9 vectors were designed to create two double-stranded breaks in each end of the gene for the NRPS-like and in one end of the NRPS-like and the other end of the putative resistance gene (to delete two genes at the same time). As repair templates, both 90 bp oligo were designed and the previously created plasmids were used. Both strains of *A. oryzae* A1560 $\Delta pyrG$, Δku and $\Delta pyrG$ were transformed with either an oligo or plasmid, but only transformants in the *A. oryzae* A1560 $\Delta pyrG$ strain were obtained. These were tested using analytical tissue PCR with multiplexing primers, however none gave the desired results. In

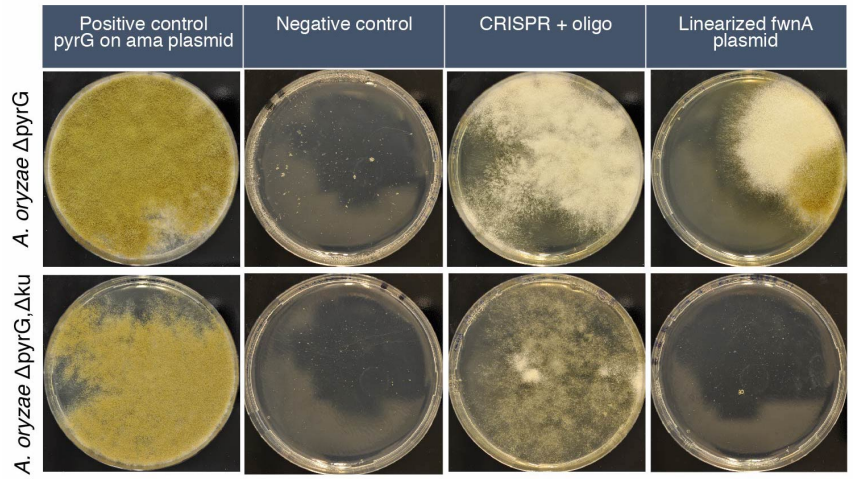


Figure 4.2 Overview of transformants of *A. oryzae*. The top row shows transformants of the *A. oryzae* A1560 Δ*pyrG* strain while the bottom row shows transformants of the *A. oryzae* A1560 Δ*pyrG,* Δ*ku*. First column positive control AMA1 (autonomous maintenance in *Aspergillus*). Second column negative control no DNA transformed. Third column, using CRISPR and 90 bp oligo as repair template. Fourth column classical gene targeting using homologous recombination – transformation with linearized plasmid.

the oligo version, the genes were not deleted and in the plasmid version *pyrG* was not inserted instead of the targeted gene.

The initial plan of deleting the cluster genes in the native host was not fulfilled. Based on the results and test made it seems that there might be a problem with the strain used *A. oryzae* A1560 Δ*pyrG,* Δ*ku* since it was very difficult to get transformants using HR, and the CRISPR based method did not give a higher efficiency as expected. Instead of trying to fix this problem, we decided to use a different strategies of heterologous expression in *A. nidulans*.

Strategy II: Heterologous expression in *A. nidulans*

Many studies of secondary metabolite gene clusters have been conducted in heterologous hosts. *A. nidulans* has been used extensively [108, 109, 110, 111, 112], and was chosen as host for heterologous expression of the *A. oryzae* cluster. An available *A. nidulans* strain (IBT 29539) NID1, which is easy to work with and

has an extensive range of genetic tools (*pyrG* and *argB* as auxotrophic marker and is deficient in non-homologous end joining (NHEJ)) [113].

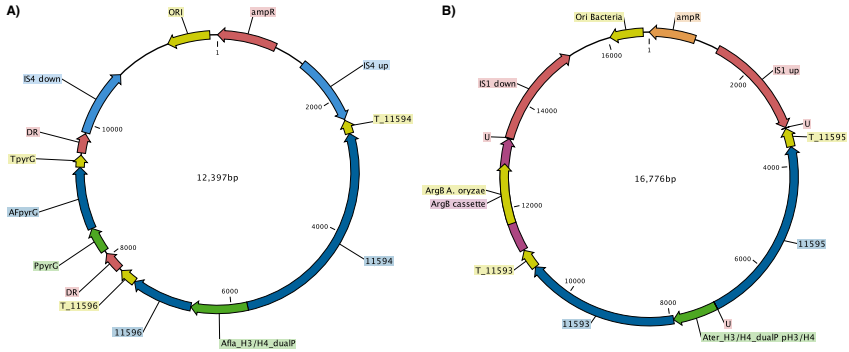


Figure 4.3 Overview of plasmids created for heterologous expression in *A. nidulans*, both plasmids contain origin of replication for propagation in *E. coli* and a gene conferring resistance to ampicillin. A) Plasmid pAC1560(short promoter)/pAC1562 targeting insertion site 4 [114] in *A. nidulans*, containing the NRPS-like gene (protein ID 11594) and the N-acetyltransferase (protein ID 11595) separated by a dual promoter originating from *A. flavus* between histone 3 and 4. The native terminators of the genes have been used by including 300-400 bp downstream of the genes. The selection marker is *pyrG* originating from *A. fumigatus*. B) Plasmid pAC1561 targeting insertion site 1 in *A. nidulans* contains the putative resistance gene (protein ID 11595) and the major facilitator (protein ID 11593) separated by a dual promoter originating from *A. terreus* between histone 3 and 4. The native terminators of the genes have been used by including 300-400 bp downstream of the genes. The selection marker is *argB* originating from *A. oryzae*.

The goal was to heterologously express all four cluster genes in *A. nidulans*. Two vectors were made: one targeting insertion site 4 [114] containing the NRPS-like and the N-acetyltransferase genes under the control of bidirectional histone promoter from *A. flavus* [115] and using *A. fumigatus pyrG* as the selection marker (pAC1560), Figure 4.3A. The other targeted insertion site 1 [114], used *A. oryzae argB* as the selection marker and contained the transporter and putative resistance gene under the control of bidirectional histone promoter from *A. terreus* [115] (pAC1561), Figure 4.3B.

Transformants with the NRPS-like and the N-acetyltransferase genes inserted

in insertion site 4 were created first, *A. nidulans* (*argB2, pyrG89, veA1, nkuA*Δ*, IS4::T11594-11594-AflaPh3/h4-11596-T11596::pyrG*. Unfortunately, it was later discovered that a too short version of the dual histone promoter from *A. flavus* had been used, which resulted in a reduction of the promoter strength in one end equalling about 10% of the original strength. The N-acetyltransferase gene was at this end and is thus only expected to be expressed 1/10th compared to the NRPS-like gene. This version with the short promoter is denoted: *A. nidulans* (*argB2, pyrG89, veA1, nkuA*Δ*, IS4::T11594-11594-AflaPh3*s/h4-11596-T11596::pyrG*. A new version was constructed with the NRPS-like and the N-acetyltransferase genes inserted in insertion site 4 and this time with the full length of the dual promoter (pAC1562).

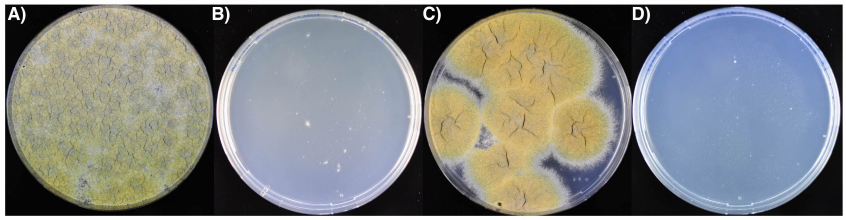


Figure 4.4 Transformations with *argB* as selection marker. The transformations were conducted in the strain *A. nidulans* (*argB2, pyrG89, veA1, nkuA*Δ*, IS4::T11594-11594-AflaPh3/h4-11596-T11596::pyrG* A) Positive control AMA1 plasmid with *argB* from *A. nidulans*. B) Control of the backbone plasmid containing only *argB* from *A. oryzae* but no cluster genes linearized for insertion with homologous recombination (HR) into IS1. C) Control of the backbone plasmid containing only *argB* from *A. nidulans*, insertion with HR in IS1. D) The plasmid with the cluster genes (11593 and 11595) under the control of the dual histone promoter from *A. terreus* with *argB* from *A. oryzae* as the selection marker linearized for insertion in IS1 using HR.

The strain with the two genes inserted and the correct dual promoter was transformed in order to insert the rest of the cluster in insertion site 1. Several attempts were made, but no transformants grew. A suspicion of the selection marker led to investigation of transformations with various methods and *argB* genes, Figure 4.4. The positive control with *argB* from *A. nidulans* as selection marker on AMA1 plasmid showed highly successful transformation with many colonies growing (Figure 4.4A). This shows that the protoplasts are competent and the selection works. Transformation of linearized plasmid containing *argB*

from *A. nidulans* for integration into insertion site 1 also showed clear colony formations (4.4C), illustrating that the method with homologous recombination and integration works. No growth was observed on both the transformation plate with linearized version of only the backbone plasmid with *argB* from *A. oryzae* (4.4B) and the transformation plate with linearized version of the same backbone plasmid with the cluster genes (4.4D), which shows that it is not the cluster genes causing the problem, but the plasmid backbone and hence most likely the selection marker *argB* from *A. oryzae*. Unfortunately we did not have enough resources to redo the construct with the *argB* from *A. nidulans* instead.

The created strains of *A. nidulans* (*argB2, pyrG89, veA1, nkuA*Δ*, IS4::T11594-11594-AflaPh3/h4-11596-T11596::pyrG* both with the short and correct promoter were used for further chemical analysis (section 4.2.2). Furthermore the created plasmids, for targeted insertion into *A. nidulans* with the cluster genes, were also used for another strategy of heterologous expression in *A. aculeatinus* elucidated below (section 4.2.2).

Strategy III: Heterologous expression in *A. aculeatinus* by random integration

Heterologous expression can be affected by the host, so the construct (pAC1560 AND pAC1562) with the NRPS-like and the N-acetyltransferase genes were also inserted randomly into *A. aculeatinus* Δ*pyrG* to investigate, if the novel compound could be found here. The vectors created for targeted insertion in *A. nidulans* were linearized and transformed randomly into the genome of *A. aculeatinus* Δ*pyrG* using *pyrG* as the selection marker, *A. aculeatinus* Δ*pyrG::11594-11596::pyrG*. Both versions with the short and the correct promoter were created and used for chemical analysis by ultra-high performance liquid chromatography – diode array detection – quadrupole time-of-flight mass spectrometry (UHPLC-DAD-QTOFMS).

Chemical analysis of the generated mutants

Firstly, chemical analysis was performed on the *A. nidulans* (*argB2, pyrG89, veA1, nkuA*Δ*, IS4::T11594-11594-AflaPh3*s/h4-11596-T11596::pyrG* and the *A. aculeatinus* Δ*pyrG:11594-11596::pyrG* strains with the short promoter. The chemical analysis of the *A. nidulans* strain was performed for the control (*A. nidulans argB2, pyrG89, veA1,* Δ*nku*IS4:pyrG) and for one mutant, while chemical analysis of the *A. aculeatinus* strains was performed for the wild type and four mutants

since the genes are integrated randomly and the expression and effect therefore might vary. The strains were grown for 7 days on MM and YES media in the dark at 37°C for *A. nidulans* and 30°C for *A. aculeatinus* and the chemical analysis was performed in triplicates.

From these strains we expected that the NRPS-like compound should be produced and potentially part of it N-acetylated. From the *A. nidulans* strain, the only significant difference in the chromatograms compared to the wild type was the production of several stress-response related compounds previously identified in other *A. nidulans* mutants (data not shown).

The hypothesis behind the selection of this cluster is that it should be bioactive producing a toxic compound and the putative resistance gene should rescue the producer. We would therefore expect an effect on the heterologous host expressing the cluster, if no resistance gene is included. The insertion of the biosynthetic cluster genes in *A. nidulans* caused the production of stress-related compounds, indicating an effect and a response in the heterologous producer. However, no new compound was detected in *A. nidulans*; one explanation for this could be, if the compound is highly active and toxic to *A. nidulans*, then very low concentrations not detected by our analytical methods could be causing the stress. This could therefore suggest that the cluster truly is bioactive.

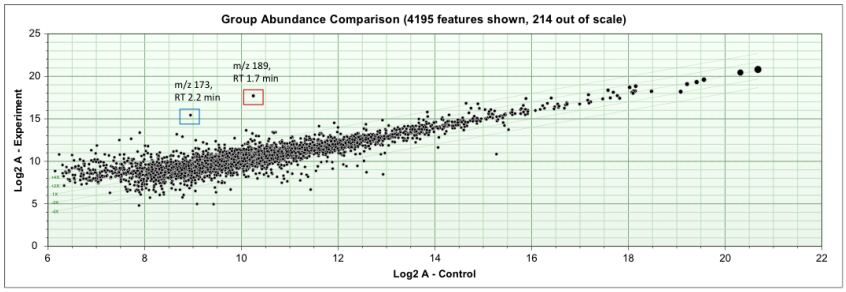


Figure 4.5 Overview of Group Abundance Comparison plot of mutants versus control from Agilent Mass profiler to show the differences and similarities in chemical profiles between the mutants *A. aculeatinus* Δ*pyrG,::11594-11596::pyrG* and the control (WT) to help identify differential compounds of interest. Two compounds were found to significantly differentiate with retention time (RT) of 1.7 and 2.2 minutes and mass to charge (m/z) for $[M+H]^+$ 189.0869 and 173.1285 respectively.

From the *A. aculeatinus* Δ*pyrG,::11594-11596::pyrG* strain with the genes in-

serted randomly, four mutants were analyzed and compared to the wild type (all in triplicates) using the program Agilent Mass Profiler, Figure 4.5. Two compounds, with mass to charge (*m/z*) for $[M+H]^+$ 189.0869 and 173.1285 and retention time (RT) 1.7 and 2.2 minutes respectively, were identified in all the mutants and not in the wild type (WT), suggesting they are related to the inserted genes. The compound at RT 1.7 min with $[M+H]^+$ 189.0869 was tentatively identified as N-acetyl-glutamine based on the HRMS and MS/HRMS data. This was verified by comparison of HRMS, MS/HRMS and retention time to the commercially available standard, appendix D Figure D.1. One possible hit for the compound with RT 2.2 minutes and $[M+H]^+$ 173.1285 was be N-acetyl-valine methylamide supported by the HRMS and MS/HRMS data, appendix D Figure D.2.

In summary, heterologous expression in *A. nidulans* revealed no new compounds which had not been identified in *A. nidulans* before, but a changed metabolite profile seen before in stressed *A. niudlans* strains. Instead, heterologous expression in *A. aculeatinus* revealed two differential compounds found in all four mutants in triplicate with the genes randomly inserted and not in the wild type. These results clearly shows that there is a difference depending on what host is chosen. It could therefore generally be a good idea, when identifying a new cluster, to test a few different hosts in order to identify the best for expression of the given cluster.

It is highly probable that the two identified compounds (*m/z* for $[M+H]^+$ 189.0869 and 173.1285 and RT 1.7 and 2.2 minutes respectively) are produced by the inserted biosynthetic genes, since they are found in all *A. aculeatinus* mutants and not in the wild type. However the relationship between the compounds, if they are steps in a biosynthetic pathway or the result of promiscuous enzymes, is not clear with the current results. It could be that one of the enzymes is not efficient and potentially needs help from anther enzyme in the conversion.

Strains with the NRPS-like and N-acetyltransferase genes and the correct, non-truncated, promoter were subsequently analyzed. Here, we expected even more compound to be acetylated as the promoter should be strong in both ends. However, the chemical analysis did not show this. For the *A. nidulans* (*argB2, pyrG89, veA1, nkuA*Δ*, IS4::11594-11596::pyrG* mutants, no significant differences were found between the mutants and control. Using the Agilent Mass Profiler, no significant differences were seen when comparing the *A. aculeatinus* Δ*pyrG*:11594-11596 mutants and the wild type, Figure 4.6A. If only looking at the features unique to the *A. aculeatinus* mutants, Figure 4.6B, a compound was identified, which was the same as the compound identified in the first experiment with RT

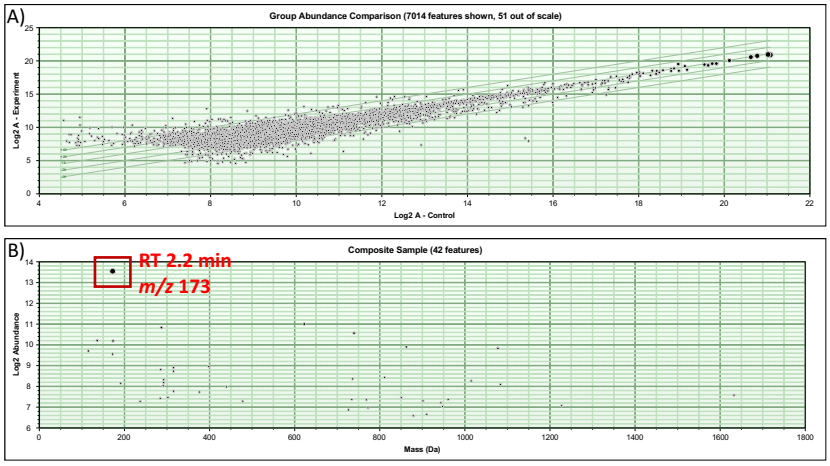


Figure 4.6 Overview from Agilent Mass Profiler showing the differences in chemical profiles between the mutants *A. aculeatinus* Δ*pyrG*:11594-11596 with the correct promoter and the wild type. A) Group Abundance Comparison plot of all the mutants (four mutants in triplicates) versus the wild type (triplicates), no significant differences were observed. B) Plot showing the features that are unique to the mutants. One feature has the retention time 2.2 minutes and *m/z* for $[M+H]^+$ 173.1285.

2.2 minutes and *m/z* for $[M+H]^+$ 173.1285, but the abundance is much lower than in the previous experiment with the short promoter, appendix D Figure D.3. The compound with RT 1.7 min *m/z* for $[M+H]^+$ 189.0869 (N-acetyl-glutamine) was not identified in these mutants.

The chemical analysis was repeated for the *A. aculeatinus* Δ*pyrG*:11594-11596::pyrG mutants with the short promoter in order to verify the previous findings and secure reproducibility. In these analyses, the abundance of the two compounds observed to differentiate in previous experiments is much lower than in the first analysis (50-75% lower in mutant 1 and 2 while only trace amounts are found in mutant 3 and 4) and a shift in the most abundant compound is seen, with more compound being produced of *m/z* for $[M+H]^+$ 173.1285 than *m/z* 189.0869 opposite to the first analysis. The decrease in abundance of the compounds observed in the repeating experiment could be due to change in the expression of the genes — for instance caused by epigenetic regulation — which might have had time to down regulate the inserted genes for the second analysis. This explanation could be

tested by expression analysis using qPCR. The observed shift from one compound to the other as the most abundant is puzzling. It is seen for both the repeated experiment with the short promoter, but also in the transformants with the long promoter. One explanation could be a slight shift in the available pre-cursors affected by the small changes in the media (which could happen since the plates are from different batches of MM) combined with promiscuous enzymes able to use different substrates. In the first analyses as well as in the repetition, *A. aculeatinus* Δ*pyrG*:11594-11596::pyrG mutant 1 had the highest production of the novel compounds compared to the other mutants, this could be due to a good insertion site or more copies inserted. The same is the case for the other mutants. As such, this experiment did not reproduce the first experiment with the truncated promoter nor the experiment with the complete promoter.

Lastly *A. oryzae* (where the investigated cluster originates from) was investigated by chemical analysis in order to see if any of the identified compounds are produced by the native host. N-acetyl glutamine was detected from *A. oryzae* grown on CYA after 3 days, but only in trace amounts and could not be reproduced. This is a small indication supporting that the compounds produced by the cluster genes have been identified, although the cluster may be regulated differently in the host, or N-acetyl glutamine is an intermediate in another pathway.

The two compounds identified in the *A. aculeatinus* mutants are produced without the presence of the resistance gene which could suggest that the compounds are not bioactive, they could potentially be intermediates which most often are not bioactive and the cluster could be part of a super cluster or more of the sourrounding genes are involoved. Another possibility is that the compound does not affect *A. aculeatinus* under the grown conditions or that the concentration is too small. From a literature search it is evident that N-acetyl-glutamine have shown to have some bioactivity, it is used as a psycostimulant and in a complex with aluminium as an anti-ulcer agent [2, 3, 4]. To test the bioactivity as an antifungal compound, one possibility could be to test the toxicity of the compound on other fungi to see if the growth is impaired and what concentration is needed to get an effect. It could also be very interesting to investigate if the mode of action is on a histidine kinase as predicted by the resistance gene.

4.2.3 Conclusion

We have used three strategies to investigate the *A. oryzae* cluster containing a putative resistance gene: I) gene deletion in the native host, II) heterologous expression in *A. nidulans*, and III) heterologous expression in *A. aculeatinus*. We

have not been able to conclusively characterize the cluster and identify the final biosynthetic compound using any of these strategies. The deletion in *A. oryzae* failed, most likely due to a problem with the background strain, heterologous expression in *A. nidulans* did not give rise to new compounds, expression in *A. aculeatinus* identified two new compounds (N-acetyl-glutamine and N-acetyl-valine methylamide) but the amount and ratio of these compounds varied. The results indicates that the cluster genes are involved in the production of N-acetyl-glutamine and N-acetyl-valine methylamide. Further analysis is needed to characterize the cluster and investigate the resistance hypothesis. One way of doing this could be to express the entire cluster in *A. aculeatinus*, which has shown to be a suitable host of this cluster, to see if the same compounds are produced and if more compound is produced if the resistance gene is present. It would also be interesting to have the resistance gene under an inducible promoter which should show whether the putative resistance gene is required for survival while the cluster is expressed. Another strategy could be to express the cluster genes in yeast both individually and in combination to characterize the biosynthesis and show whether the resistance gene is required in this host.

4.2.4 Methods

Strains and media

A list of all strains used and constructed in this study is provided in supplementary Table D.1. Gene deletion in *A. oryzae* was performed using *A. oryzae* A1560 Δ*pyrG,* Δ*ku* and *A. oryzae* A1560 Δ*pyrG*. Heterologous expression was performed in *A. nidulans argB2, pyrG89, veA1,* Δ*nku* (NID1), while heterologous expression in *A. aculeatinus* was performed in a Δ*pyrG* strain. *Escherichia coli* strain DH5*alpha* was used for plasmid propagation. The background strains of *A. oryzae* A1560, *A. nidulans* and *A. aculeatinus* were obtained from the in house strain collection.

Aspergillus solid and liquid minimal medium (MM) and transformation medium (TM) were prepared as described by Nødvig et al. [104] and yeast extract sucrose (YES) medium was prepared as described by Samson et al. [116]. The medium was supplemented with 10 mM uridine and 10 mM uracil or 4 mM arginine when necessary. *E. coli* was grown on solid and liquid Luria-Betrani (LB) medium containing 10 g/l tryptone (Bacto), 5 g/L yeast extract (Bacto), 10 g/l NaCL (pH 7.0) supplemented with 100 μg/ml ampicillin.

Vector and strain construction

All DNA fragments used for cloning were amplified in 35 cycles using PfuX polymerase [117] and touch-down PCR programme with the annealing temperature from 65-56°C. The standard reaction used included 200 μM dNTPs, 1x Phusion HF buffer (New England Biolabs, USA), 0.4 μM primers (purchased from Integrated DNA technologies, Belgium), 1 U PfuX7, $\leq$ 10 ng gDNA with a total volume of 50 μl. The primers used for amplification are listed in Supplementary Table D.2. Genomic DNA (gDNA) was extracted from *A. oryzae* A1560 using the FastDNA SPIN Kit for soil DNA extraction (MP Biomedicals, USA) and used as template with $\leq$ 10 ng/μl.

Plasmids were constructed using Uracil-Specific Excision Reagent (USER) cloning or fusion [118], Supplementary Table D.3. All constructed plasmids contained an origin of replication for propagation in *E. coli* and the ampicillin resistance gene. Plasmids for gene deletion were constructed by introducing the PCR-amplified up and downstream fragments into two specific USER casettes on each side of the *A. flavus pyrG* selection marker.

Plasmids were purified using the GenElute Plasmid Miniprep Kit (Sigma-Aldrich) and were linearized with *SwaI/NotI* (New England Biolabs) prior to transformation when using homologous recombination.

Protoplastation and transformation

In order to make protoplasts, spores from one or two agar plates were harvested and filtered through sterile Miracloth (EMD Milipore) into a shakeflask with 100 mL of appropriate media with supplements (MM for *A. nidulans* and YDP for *A. oryzae*). The shakeflasks were incubated at appropriate temperature (37°C for *A. nidulans* and 30°C for *A. oryzae*) for 4-5 hours (the spores should just have started to germinate, so when looking in the microscope a few have started to "bud", while the majority still looks like spores). The culture was then spun down at 5,000g for 10 min, and the medium was discarded. Subsequently the spores in the pellet were digested in a 20 ml sterile solution of 40 mg/ml Glucanex (Novozymes) dissolved in APB. The spore Glucanex solution was incubated for digestion at 30°C 150 rpm for 1-1.5 hour. Following digestion the solution was spun down at 1500 g for 10 min and the APB supernatant solution was removed. The spores were washed 1-2 times in ST buffer (1 M sorbitol; 50mM Tris; pH 8) and spun at 1500 g for 10 min each time. Finally the washed spores were resuspended in 2-5 ml STC (1 M sorbitol; 50mM Tris; 50mM $CaCl_2$; pH 8) for a final concentration of 1.2×10^7.

The protoplasts were stored at -80°C.

Transformation was performed as described by Nødvig et al. [104]. All candidate transformants were purified by streaking on agar plates prior to verification. All strains were verified by tissue-PCR analysis using mycelium as the source of DNA and including wild type gDNA as control. The primers can be found in Supplementary Table D.2. Primers for gene- deletion analysis were designed to bind up- and downstream outside the region eliminated by the gene-targeting substrate. Primers for gene insertions were designed to bind up and downstream of the insertion site and in each end of the inserting genes, four primers were included in one reaction giving one large band if the insertion was unsuccessful and two smaller bands if the insertion is successful.

Chemical analysis

The mutants and wild types were grown on MM and YES agar plates for 7 days, subsequently 5 plugs were taken across the colony, 800 μl isopropanol:ethyl acetate (1:3 v/v) with 1% formic acid was added and ultrasonicated for 1 hour. The liquid sample was transferred to another tube and evaporated, following 300 μL methanol was added to redissolve the pellets and the samples were ultrasonicated 20 minutes. Samples were then centrifuged at max g-power for 2-3 minutes, and afterwards 150 μL of the supernatant was transferred to HPLC vials.

Ultra-high performance liquid chromatography – diode array detection – quadrupole time-of-flight mass spectrometry (UHPLC-DAD-QTOFMS) was performed on an Agilent Infinity 1290 UHPLC system equipped with a diode array detector. Separation was obtained on a 250 × 2.1 mm i.d., 2.7 μm, Poroshell 120 Phenyl Hexyl column (Agilent Technologies, Santa Clara, CA) held at 60 °C. The sample, 1 μL, was eluted with a flow rate of 0.35 mL/min using A: a linear gradient 10% acetonitrile in Mili-Q water buffered with 20 mM formic acid increased to 100% in 15 min., staying there for 2 min. before returning to 10% in 0.1 min., held for 3 min. before the following run. Mass spectrometry (MS) detection was performed on an Agilent 6545 QTOF MS equipped with an Agilent dual jet stream electro spray ion (ESI) source with a drying gas temperature of 160 °C, gas flow of 13 L/min, sheath gas temperature of 300 °C, and flow of 16 L/min. Capillary voltage was set to 4000 V, and nozzle voltage, to 500 V in positive mode. MS spectra were recorded as centroid data, at an m/z of 100 to 1,700, and auto MS/HRMS fragmentation was performed at three collisions energies (10, 20, 40 eV), on the three most intense precursor peaks per cycle. The acquisition was 10 spectra/s. Data was handled in the software Agilent MassHunter Mass Profiler and MassHunter

Qualitative Analysis (Agilent Technologies, Santa Clara, CA).

5 Conclusion

With the increasing number of fungal whole genome sequences, exciting opportunities emerge but also challenges. Both of which are addressed in this thesis. The opportunities include gaining new insights and creating knowledge-based hypothesis, whereas the main challenges are how to use the data, what to focus on with the unprecedented amount of data, and — post-data analysis — which projects to initiate based on the data.

An important discovery from the first *Aspergillus* whole genome sequences was the large potential of secondary metabolite biosynthesis. This initiated opportunities in genome mining, but it still remains a challenge to link biosynthetic gene clusters to the compounds they produce. I have addressed this challenge both in terms of presenting state-of-the-art in Chapter 2 and by predicting biosynthetic gene clusters for aflatoxin, chlorflavonin, novofumigatonin and ochrindol based on the genome sequences in Section 3.1, thereby addressing aim 1. Chapter 3 also addresses two other aims, publishing genomes and investigating the biological and chemical diversity within *Aspergillus* species (aim 2 and 3). With Paper I and Manuscript II, we will publish 6 and 19 genomes respectively, moreover we have characterized the genomes and used them to investigate the diversity. In Manuscript II (section 3.2, I focused on the important *Flavi* section. With genome sequences covering the entire section, we had a unique opportunity to explore the diversity and similarities across the *Flavi* section. From this work we have shown that section *Flavi* has a very high potential for secondary metabolite production and carbohydrate active enzymes. We have, through phylogenetic analyses, shown differences in the evolutionary pattern compared to the common theories, and from synteny analysis shown potential patterns of genome evolution with highly conserved regions and variable blocks.

The first whole genome sequencing data of *Aspergillus* species revealed a large number of predicted secondary metabolite gene clusters. With the current molecular methods, it is unfeasible to investigate all the predicted clusters in the search for novel bioactive compounds, which is why a targeted method is needed to identify which clusters to elucidate. We have addressed this need using a hypothesis-driven approach based on duplicated self-resistance genes and comparative genomics, presented in Chapter 4, Manuscript III. We chose to attack the challenge by develop-

ing a pipeline identifying putative resistance genes in biosynthetic gene clusters. Applying the pipeline to 51 *Aspergillus* and *Penicillium* genomes, we were able to validate the method by identification of a known cluster with a verified self-resistance gene as well as the identification of 71 other protein families containing unverified resistance genes. We have investigated a selected cluster with a putative resistance gene in an attempt to verify the identified cluster and with a hope of finding a novel bioactive compound. The efforts have so far only produced inconclusive results. In the future the pipeline or the rationale behind it can be used to direct experimental efforts in natural product discovery in the quest to identify novel anti-fungals and meet the desperate need.

The work presented in this thesis will function as a reference for future work both within research in section *Flavi* and within the field of linking compounds and gene clusters. Our work has provided reference genomes and an overview of the potential of section *Flavi* and other researchers can build on this and create new insights. We have shown various ways of identifying biosynthetic gene clusters from compounds which hopefully will inspire other researchers to use the strategies and resources to link their favorite cluster genes and compounds. Finally, we have provided the natural product discovery field with a new targeted method of identifying likely bioactive compounds and gene clusters. The identified resistance genes, targets and clusters will hopefully be examined and tested in the near future to provide much needed new antifungals.

In the future we are likely to see an increasing number of studies using comparative genomics to create knowledge-based hypothesis. Genome sequencing projects such as the *Aspergillus* whole genus project and the 1000 fungal genomes provide an enormous resource to the scientific community not seen before within the fungal field, opening for exciting opportunities in a variety of fields including basic research in the evolution and adaptation and applied research in metabolic engineering, and enzyme and drug discovery.

Bibliography

[1] Yudai Matsuda et al. “Novofumigatonin biosynthesis involves a non-heme iron-dependent endoperoxide isomerase for orthoester formation”. In: *Nature Communications* 9.1 (2018), p. 2587. ISSN: 20411723.

[2] H. Tanaka, K Shuto, and H Marumo. “Effect of N-acetyl-L-glutamine aluminum complex (KW-110), an antiulcer agent, on the non-steroidal anti-inflammatory drug-induced exacerbation of gastric ulcer in rats.” In: *Japanese journal of pharmacology* 32.2 (1982), pp. 307–13. ISSN: 0021-5198.

[3] J. Elks. *The Dictionary of Drugs - Chemical Data, Structures and bibliographies*. Ed. by J. Elks and C.R. Ganellin. Springer US, 2014, p. 3. ISBN: 9781475720853.

[4] Rui Zhang et al. “Neuroprotective effects of Aceglutamide on motor function in a rat model of cerebral ischemia and reperfusion”. In: *Restorative Neurology and Neuroscience* 33.5 (2015), pp. 741–759. ISSN: 18783627.

[5] Leho Tedersoo et al. “Global diversity and geography of soil fungi”. In: *Science* 346.6213 (Aug. 2014), p. 1256688. ISSN: 10959203.

[6] Meredith Blackwell. “The Fungi: 1, 2, 3 ... 5.1 million species?” In: *American Journal of Botany* 98.3 (2011), pp. 426–438.

[7] Elizabeth Pennisi. “Armed and dangerous.” In: *Science (New York, N.Y.)* 327.5967 (Feb. 2010), pp. 804–5. ISSN: 1095-9203.

[8] Matthew C. Fisher et al. “Emerging fungal threats to animal, plant and ecosystem health”. In: *Nature* 484.7393 (2012), pp. 186–194. ISSN: 00280836. arXiv: NIHMS150003.

[9] Gordon D Brown, David W Denning, Neil A R Gow, Stuart M Levitz, Mihai G Netea, and Theodore C White. “Hidden killers: human fungal infections.” In: *Science translational medicine* 4.165 (2012), 165rv13. ISSN: 1946-6242. arXiv: arXiv:1011.1669v3.

[10] D. Kontoyiannis and G. Bodey. “Invasive aspergillosis in 2002: An update”. In: *European Journal of Clinical Microbiology and Infectious Diseases* 21.3 (2002), pp. 161–172. ISSN: 09349723.

[11] Peter G. Pappas et al. “Invasive Fungal Infections among Organ Transplant Recipients: Results of the Transplant-Associated Infection Surveillance Network (TRANSNET)”. In: *Clinical Infectious Diseases* 50.8 (2010), pp. 1101–1111. ISSN: 1058-4838.

[12] Dimitrios P. Kontoyiannis et al. “Prospective Surveillance for Invasive Fungal Infections in Hematopoietic Stem Cell Transplant Recipients, 2001–2006: Overview of the Transplant-Associated Infection Surveillance Network (TRANSNET) Database”. In: *Clinical Infectious Diseases* 50.8 (2010), pp. 1091–1100. ISSN: 1058-4838.

[13] Frank C. Odds, Alistair J.P. Brown, and Neil A.R. Gow. "Antifungal agents: Mechanisms of action". In: *Trends in Microbiology* 11.6 (2003), pp. 272–279. ISSN: 0966842X.

[14] Paul Taylor. *Antifungal Drugs: Technologies and Global Markets*. Tech. rep. BCC Research Phamaceutical Report, 2017, PHM029F.

[15] Susan J. Howard et al. "Frequency and evolution of azole resistance in Aspergillus fumigatus associated with treatment failure". In: *Emerging Infectious Diseases* 15.7 (2009), pp. 1068–1076. ISSN: 10806040.

[16] Franqueline Reichert-Lima et al. "Surveillance for azoles resistance in Aspergillus spp. highlights a high number of amphotericin B-resistant isolates". In: *Mycoses* 61.6 (June 2018), pp. 360–365. ISSN: 14390507.

[17] Eugénia Pinto et al. "Aspergillus species and antifungals susceptibility in clinical setting in the north of Portugal: Cryptic species and emerging azoles resistance in A. fumigatus". In: *Frontiers in Microbiology* 9 (2018), p. 1656. ISSN: 1664302X.

[18] Paul E. Verweij, Eveline Snelders, Gert HJ Kema, Emilia Mellado, and Willem JG Melchers. *Azole resistance in Aspergillus fumigatus: a side-effect of environmental fungicide use?* Tech. rep. 12. 2009, pp. 789–795.

[19] Klaus Leth Mortensen, Emilia Mellado, Cornelia Lass-Flörl, Juan Luis Rodriguez-Tudela, Helle Krogh Johansen, and Maiken Cavling Arendrup. "Environmental study of azole-resistant Aspergillus fumigatus and other aspergilli in Austria, Denmark, and Spain". In: *Antimicrobial Agents and Chemotherapy* 54.11 (2010), pp. 4545–4549. ISSN: 00664804.

[20] Jens C. Frisvad and Thomas O. Larsen. "Chemodiversity in the genus Aspergillus". In: *Applied Microbiology and Biotechnology* 99.19 (2015), pp. 7859–7877. ISSN: 14320614. arXiv: arXiv:1011.1669v3.

[21] Henk van Liempt, Hans von Döhren, and Horst Kleinkauf. "Delta-(L-Alpha-Aminoadipyl)-L-Cysteinyl-D-Valine Synthetase from Aspergillus-Nidulans - the 1st Enzyme in Penicillin Biosynthesis Is a Multifunctional Peptide Synthetase". In: *Journal of Biological Chemistry* 264.7 (1989), pp. 3680–3684.

[22] Lee Hendrickson et al. "Lovastatin biosynthesis in Aspergillus terreus: Characterization of blocked mutants, enzyme activities and a multifunctional polyketide synthase gene". In: *Chemistry and Biology* 6.7 (1999), pp. 429–439. ISSN: 10745521.

[23] James E. Galagan et al. "Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae." en. In: *Nature* 438.7071 (Dec. 2005), pp. 1105–15. ISSN: 1476-4687.

[24] Masayuki Machida et al. "Genome sequencing and analysis of Aspergillus oryzae". In: *Nature* 438 (2005), pp. 1157–1161.

[25] William C Nierman et al. "Genomic sequence of the pathogenic and allergenic filamentous fungus Aspergillus fumigatus". In: *Nature* 438 (2005), pp. 1151–1156.

[26] Nancy P. Keller, Geoffrey Turner, and Joan W. Bennett. "Fungal secondary metabolism – from biochemistry to genomics". In: *Nature Reviews Microbiology* 3.12 (2005), pp. 937–947.

[27] R. A. Samson et al. "Phylogeny,identification and nomenclature of the genus Aspergillus". In: *Studies in Mycology* 78 (2014), pp. 141–173. ISSN: 0166-0616.

[28] S. Kocsubé et al. "Aspergillus is monophyletic: Evidence from multiple gene phylogenies and extrolites profiles". In: *Studies in Mycology* 85 (2016), pp. 199–213. ISSN: 01660616.

[29] Jos Houbraken, Ronald P. de Vries, and Robert A. Samson. "Modern taxonomy of biotechnologically important Aspergillus and Penicillium species". In: *Advances in Applied Microbiology*. Vol. 86. 2014, pp. 199–249. ISBN: 9780128002629.

[30] Vit Hubka, Alena Nováková, Miroslav Kolařík, Željko Jurjević, and Stephen W. Peterson. "Revision of ¡i¿Aspergillus¡/i¿ section ¡i¿Flavipedes¡/i¿ : seven new species and proposal of section ¡i¿Jani¡/i¿ sect. nov." In: *Mycologia* 107.1 (2015), pp. 169–208. ISSN: 0027-5514.

[31] Jean-paul Latgé. "Aspergillus fumigatus and Aspergillosis Aspergillus fumigatus and Aspergillosis". In: *Clinical Microbiology Reviews* 12.2 (1999), pp. 310–350.

[32] M. A. Pfaller, P. G. Pappas, and J. R. Wingard. *Invasive Fungal Pathogens: Current Epidemiological Trends*. Tech. rep. Supplement 1. 2006, S3–S14.

[33] Helioswilton Sales-Campos, Ludmilla Tonani, Cristina Ribeiro, Barros Cardoso, Márcia Regina, and Von Zeska Kress. "The Immune Interplay between the Host and the Pathogen in Aspergillus fumigatus Lung Infection". In: *BioMed Research International* (2013), p. ID 693023.

[34] Jean-Paul Latgé. "The pathobiology of Aspergillus fumigatus". In: *TRENDS in Microbiology* 9.8 (2001), pp. 382–389.

[35] János Varga, Nikolett Baranyi, Muthusamy Chandrasekaran, Csaba Vágvölgyi, and Sándor Kocsubé. *Mycotoxin producers in the Aspergillus genus: An update*. Tech. rep. 2. 2015, pp. 151–167.

[36] Thomas W Kensler, Bill D Roebuck, Gerald N Wogan, and John D Groopman. "Aflatoxin: A 50-Year Odyssey of Mechanistic and Translational Toxicology". In: *TOXICOLOGICAL SCIENCES* 120.S1 (2011), pp. 28–48.

[37] Pradeep Kumar, Dipendra K. Mahato, Madhu Kamle, Tapan K. Mohanta, and Sang G. Kang. "Aflatoxins: A global concern for food safety, human health and their management". In: *Frontiers in Microbiology* 7 (2017), p. 2170. ISSN: 1664302X.

[38] E. Schuster, N. Dunn-Coleman, J. Frisvad, and P. Van Dijck. "On the safety of Aspergillus niger - A review". In: *Applied Microbiology and Biotechnology* 59.4-5 (2002), pp. 426–435. ISSN: 01757598.

[39] Marin Berovic and Matic Legisa. "Citric acid production". In: *Biotechnology Annual Review* 13 (2007), pp. 303–343. ISSN: 13872656.

[40] Belén Max, José Manuel Salgado, Noelia Rodríguez, Sandra Cortés, Attilio Converti, and José Manuel Domínguez. "Biotechnological production of citric acid." In: *Brazilian journal of microbiology : [publication of the Brazilian Society for Microbiology]* 41.4 (2010), pp. 862–75. ISSN: 1517-8382.

[41] David Lubertozzi and Jay D. Keasling. "Developing Aspergillus as a host for heterologous expression". In: *Biotechnology Advances* 27.1 (2009), pp. 53–75. ISSN: 07349750.

[42] André Fleiner and Petra Dersch. "Expression and export: Recombinant protein production systems for Aspergillus". In: *Applied Microbiology and Biotechnology* 87.4 (2010), pp. 1255–1270. ISSN: 01757598.

[43] Satoshi Wakai, Takayoshi Arazoe, Chiaki Ogino, and Akihiko Kondo. "Future insights in fungal metabolic engineering". In: *Bioresource Technology* 245 (2017), pp. 1314–1326. ISSN: 18732976.

[44] Tetsuo Kobayashi et al. "Genomics of Aspergillus oryzae". In: *Bioscience, Biotechnology, and Biochemistry* 71.3 (2007), pp. 646–670.

[45] Masayuki Machida, Osamu Yamada, and Katsuya Gomi. "Genomics of Aspergillus oryzae: Learning from the History of Koji Mold and Exploration of Its Future". In: *DNA Research* 15 (2008), pp. 173–183.

[46] Atsushi Sato et al. "Draft Genome Sequencing and Comparative Analysis of Aspergillus sojae NBRC4239". In: *DNA Research* 18 (2011), pp. 165–176.

[47] Helga David, Ilknur Ş Özçelik, Gerald Hofmann, and Jens Nielsen. "Analysis of Aspergillus nidulans metabolism at the genome-scale". In: *BMC Genomics* 9 (2008), p. 163. ISSN: 14712164.

[48] Marta Pecyna and Marcin Bizukojc. "Lovastatin biosynthesis by Aspergillus terreus with the simultaneous use of lactose and glycerol in a discontinuous fed-batch culture". In: *Journal of Biotechnology* 151.1 (2011), pp. 77–86. ISSN: 01681656.

[49] Diane O Inglis et al. "Comprehensive annotation of secondary metabolite biosynthetic genes and gene clusters of Aspergillus nidulans, A. fumigatus, A. niger and A. oryzae." In: *BMC microbiology* 13 (2013), p. 91. ISSN: 1471-2180.

[50] Jens C. Frisvad. "Taxonomy, chemodiversity and chemoconsistency of Aspergillus, Penicillium and Talaromyces species". In: *Frontiers in Microbiology* 5 (2015), p. 773. ISSN: 1664302X.

[51] G. A. Payne et al. "Whole genome comparison of Aspergillus flavus and A. oryzae". In: *Medical Mycology* 44 (2006), pp. 9–11.

[52] Herman J. Pel et al. "Genome sequencing and analysis of the versatile cell factory Aspergillus niger CBS 513.88". In: *Nature Biotechnology* 25.2 (2007), pp. 221–231. ISSN: 10870156.

[53] Natalie D Fedorova et al. "Genomic islands in the pathogenic filamentous fungus Aspergillus fumigatus." In: *PLoS genetics* 4.4 (2008), e1000046. ISSN: 1553-7404.

[54] Mikael R. Andersen et al. “Comparative genomics of citric-acid-producing Aspergillus niger ATCC 1015 versus enzyme-producing CBS 513.88”. In: *Genome Research* 21.6 (2011), pp. 885–897. ISSN: 10889051.

[55] James F. Sanchez, Amber D. Somoza, Nancy P. Keller, and Clay C.C. Wang. “Advances in Aspergillus secondary metabolite research in the post-genomic era”. In: *Natural Product Reports* 29.3 (2012), pp. 351–371. ISSN: 02650568. arXiv: 15334406.

[56] A. Rokas et al. “What can comparative genomics tell us about species concepts in the genus Aspergillus?” In: *Studies in Mycology* 59 (2007), pp. 11–17. ISSN: 01660616.

[57] John G. Gibbons and Antonis Rokas. “The function and evolution of the Aspergillus genome”. In: *Trends in Microbiology* 21.1 (2013), pp. 14–22. ISSN: 0966842X. arXiv: NIHMS150003.

[58] Abigail L. Lind et al. “Drivers of genetic diversity in secondary metabolic gene clusters within a fungal species”. In: *PLoS Biology* 15.11 (2017), e2003583. ISSN: 15457885.

[59] de Vries et al. “Comparative genomics reveals high biological diversity and specific adaptations in the industrially and medically important fungal genus Aspergillus”. In: *Genome Biology* 18.28 (2017).

[60] Kirstin Scherlach and Christian Hertweck. “Discovery of aspoquinolones A-D, prenylated quinoline-2-one alkaloids from Aspergillus nidulans, motivated by genome mining”. In: *Organic and Biomolecular Chemistry* 4.18 (2006), pp. 3517–3520. ISSN: 14770520.

[61] Nancy P. Keller and Thomas M. Hohn. “Metabolic pathway gene clusters in filamentous fungi”. In: *Fungal Genetics and Biology* 21.1 (1997), pp. 17–29. ISSN: 10871845.

[62] Anne Osbourn. “Secondary metabolic gene clusters: evolutionary toolkits for chemical innovation”. In: *Trends in Genetics* 26.10 (2010), pp. 449–457. ISSN: 01689525.

[63] Nancy P. Keller. “Translating biosynthetic gene clusters into fungal armor and weaponry”. In: *Nature Chemical Biology* 11.9 (2015), pp. 671–677. ISSN: 15524469.

[64] Jae-Hyuk Yu and Nancy Keller. “Regulation of Secondary Metabolism in Filamentous Fungi”. In: *Annual Review of Phytopathology* 43.1 (2005), pp. 437–458. ISSN: 0066-4286.

[65] Dirk Hoffmeister and Nancy P Keller. “Natural products of filamentous fungi: enzymes, genes, and their regulation.” In: *Natural product reports* 24.2 (2007), pp. 393–416. ISSN: 0265-0568.

[66] Nora Khaldi et al. “SMURF: Genomic mapping of fungal secondary metabolite clusters.” In: *Fungal genetics and biology : FG & B* 47.9 (Sept. 2010), pp. 736–41. ISSN: 1096-0937.

[67] David A. Hopwood. “Genetic Contributions to Understanding Polyketide Synthases”. In: *Chemical Reviews* 97.7 (1997), pp. 2465–2498. ISSN: 0009-2665.

[68] Russell J. Cox and Thomas J. Simpson. “Chapter 3 Fungal Type I Polyketide Synthases”. In: *Methods in Enzymology*. Vol. 459. 09. 2009, pp. 49–78. ISBN: 9780123745910.

[69] Kira J. Weissman. “The structural biology of biosynthetic megaenzymes”. In: *Nature Chemical Biology* 11.9 (2015), pp. 660–670. ISSN: 15524469.

[70] Mohamed A. Marahiel. “A structural model for multimodular NRPS assembly lines”. In: *Natural Product Reports* 33.2 (2016), pp. 136–140. ISSN: 14604752.

[71] Jonathan Kennedy, Karine Auclair, Steven G. Kendrew, Cheonseok Park, John C. Vederas, and C. Richard Hutchinson. “Modulation of polyketide synthase activity by accessory proteins during lovastatin biosynthesis”. In: *Science* 284.5418 (1999), pp. 1368–1372. ISSN: 00368075.

[72] Rachel V.K. Cochrane et al. “Production of New Cladosporin Analogues by Reconstitution of the Polyketide Synthases Responsible for the Biosynthesis of this Antimalarial Agent”. In: *Angewandte Chemie - International Edition* 55.2 (Jan. 2016), pp. 664–668. ISSN: 15213773.

[73] Gerald Bills, Yan Li, Li Chen, Qun Yue, Xue Mei Niu, and Zhiqiang An. “New insights into the echinocandins and other fungal non-ribosomal peptides and peptaibiotics”. In: *Natural Product Reports* 31.10 (2014), pp. 1348–1375. ISSN: 14604752.

[74] Kathryn E. Bushley et al. “The Genome of Tolypocladium inflatum: Evolution, Organization, and Expression of the Cyclosporin Biosynthetic Gene Cluster”. In: *PLoS Genetics* 9.6 (2013), p. 1003496. ISSN: 15537390.

[75] Arnold L. Demain. “HOW DO ANTIBIOTIC-PRODUCING MICROORGANISMS AVOID SUICIDE?” In: *Annals of the New York Academy of Sciences* 235.1 (May 1974), pp. 601–612. ISSN: 0077-8923.

[76] Eric Cundliff. *How antibiotic-producing organisms avoid suicide*. Tech. rep. 1989, pp. 207–233.

[77] Daniel H. Scharf, Nicole Remme, Thorsten Heinekamp, Peter Hortschansky, Axel A. Brakhage, and Christian Hertweck. “Transannular disulfide formation in gliotoxin biosynthesis and its role in self-resistance of the human pathogen Aspergillus fumigatus”. In: *Journal of the American Chemical Society* 132.29 (2010), pp. 10136–10141. ISSN: 00027863.

[78] Stephen K. Dolan, Grainne O’Keeffe, Gary W. Jones, and Sean Doyle. “Resistance is not futile: Gliotoxin biosynthesis, functionality and utility”. In: *Trends in Microbiology* 23.7 (2015), pp. 419–428. ISSN: 18784380.

[79] N. J. Alexander, S. P. McCormick, and T. M. Hohn. “TRI12, a trichothecene efflux pump from Fusarium sporotrichioides: Gene isolation and expression in yeast”. In: *Molecular and General Genetics* 261.6 (1999), pp. 977–984. ISSN: 00268925.

[80] Jon Menke, Yanhong Dong, and H. Corby Kistler. "¡i¿Fusarium graminearum¡/i¿ Tri12p Influences Virulence to Wheat and Trichothecene Accumulation". In: *Molecular Plant-Microbe Interactions* 25.11 (2012), pp. 1408–1418. ISSN: 0894-0282.

[81] Michael H. Perlin, Jared Andrews, and Su San Toh. "Essential letters in the fungal alphabet: ABC and MFS transporters and their roles in survival and pathogenicity". In: *Advances in Genetics*. Vol. 85. 2014, pp. 201–253. ISBN: 9780128002711.

[82] H. Corby Kistler and Karen Broz. "Cellular compartmentalization of secondary metabolism". In: *Frontiers in Microbiology* 6 (Feb. 2015), p. 68. ISSN: 1664302X.

[83] A. Chanda et al. *A key role for vesicles in fungal secondary metabolism*. Tech. rep. 46. 2009, pp. 19533–19538.

[84] Ludmila V. Roze, Anindya Chanda, and John E. Linz. "Compartmentalization and molecular traffic in secondary metabolism: A new understanding of established cellular processes". In: *Fungal Genetics and Biology* 48.1 (2011), pp. 35–48. ISSN: 10871845.

[85] Y. Abe et al. "Effect of increased dosage of the ML-236B (compactin) biosynthetic gene cluster on ML-236B production in Penicillium citrinum". In: *Molecular Genetics and Genomics* 268.1 (2002), pp. 130–137. ISSN: 16174615.

[86] P. Wiemann, C.-J. Guo, J. M. Palmer, R. Sekonyela, C. C. C. Wang, and N. P. Keller. "Prototype of an intertwined secondary-metabolite supercluster". In: *Proceedings of the National Academy of Sciences* 110.42 (2013), pp. 17065–17070. ISSN: 0027-8424.

[87] Torsten Bak Regueira, Kanchana Rueksomtawin Kildegaard, Bjarne Gram Hansen, Uffe H. Mortensen, Christian Hertweck, and Jens Nielsen. "Molecular basis for mycophenolic acid biosynthesis in Penicillium brevicompactum". In: *Applied and Environmental Microbiology* 77.9 (2011), pp. 3035–3043. ISSN: 00992240.

[88] Xin E Sun, Bjarne Gram Hansen, and Lizbeth Hedstrom. "Kinetically Controlled Drug Resistance". In: *Journal of Biological Chemistry* 286.47 (2011), pp. 40595–40600. ISSN: 1083351x, 00219258.

[89] Bjarne G Hansen et al. "A new class of IMP dehydrogenase with a role in self-resistance of mycophenolic acid producing fungi." In: *BMC microbiology* 11 (Jan. 2011), p. 202. ISSN: 1471-2180.

[90] Hsu-Hua Yeh et al. "Resistance gene-guided genome mining: Serial promoter exchanges in Aspergillus nidulans reveal the biosynthetic pathway for fellutamide B, a proteasome inhibitor". In: *ACS Chemical Biology* 11.8 (2016), pp. 2275–2284. ISSN: 1554-8929.

[91] Yan Yan et al. "Resistance-gene-directed discovery of a natural-product herbicide with a new mode of action". In: *Nature* 559.7714 (2018), pp. 415–418. ISSN: 14764687.

[92] Tammi C. Vesth et al. "Investigation of inter- and intra-species variation through genome sequencing of Aspergillus section Nigri". In: *Nature Genetics* Accepted - (2018).

[93] Inge Kjærbølling et al. “Linking secondary metabolites to gene clusters through genome sequencing of six diverse Aspergillus species”. In: *Proceedings of the National Academy of Sciences* 115.4 (2018), E753–E761. ISSN: 0027-8424.

[94] Amjad Ali et al. “Microbial Comparative Genomics: An Overview of Tools and Insights Into The Genus Corynebacterium”. In: *Journal of Bacteriology & Parasitology* 04.2 (2013), p. 1000167. ISSN: 21559597.

[95] J. Yu et al. “A review of bioinformatics platforms for comparative genomics. Recent developments of the EDGAR 2.0 platform and its utility for taxonomic and phylogenetic studies”. In: *Journal of Biotechnology* 261 (Nov. 2017), pp. 2–9. ISSN: 18734863.

[96] Jane Lind Nybo Rasmussen, Sebastian Theobald, Julian Brandl, Tammi Camilla Vesth, and Mikael Rørdam Andersen. “Approaches for Comparative Genomics in Aspergillus and Penicillium”. In: *Aspergillus and Penicillium in the post-genomic era*. Ed. by Ronald P. de Vries, Isabelle B. Gelber, and Mikael R. Andersen. Caister Academic Press, 2016. Chap. Chapter 4, pp. 43–73.

[97] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. “Basic Local Alignment Search Tool”. In: *Journal of molecular biology* 215.3 (1990), pp. 403–410.

[98] The Gene Ontology Consortium et al. “Gene Ontology: tool for the unification of biology”. In: *Nature Genetics* 25 (2000), pp. 25–29.

[99] The Gene Ontology Consortium. “Expansion of the Gene Ontology knowledgebase and resources”. In: *Nucleic Acids Research* 45 (2017), pp. D331–D338.

[100] Roman L Tatusov et al. “The COG database: an updated version includes eukaryotes”. In: *BMC Bioinformatics* 4 (2003), p. 41.

[101] Philip Jones et al. “InterProScan 5: Genome-scale protein function classification”. In: *Bioinformatics* 30.9 (2014), pp. 1236–1240. ISSN: 14602059.

[102] Robert D. Finn et al. “InterPro in 2017-beyond protein family and domain annotations”. In: *Nucleic Acids Research* 45.D1 (2017), pp. D190–D199. ISSN: 13624962.

[103] Kai Blin et al. “AntiSMASH 4.0 - improvements in chemistry prediction and gene cluster boundary identification”. In: *Nucleic Acids Research* 45.Web server issue (2017), W36–W41. ISSN: 13624962.

[104] Christina S. Nødvig, Jakob B. Nielsen, Martin E. Kogle, and Uffe H. Mortensen. “A CRISPR-Cas9 system for genetic engineering of filamentous fungi”. In: *PLoS ONE* 10.7 (2015), e0133085. ISSN: 19326203. arXiv: arXiv:1011.1669v3.

[105] Christina S. Nødvig et al. “Efficient oligo nucleotide mediated CRISPR-Cas9 gene editing in Aspergilli”. In: *Fungal Genetics and Biology* 115 (June 2018), pp. 78–89. ISSN: 10960937.

[106] M. R. Andersen, W. Vongsangnak, G. Panagiotou, M. P. Salazar, L. Lehmann, and J. Nielsen. "A trispecies Aspergillus microarray: Comparative transcriptomics of three Aspergillus species". In: *Proceedings of the National Academy of Sciences* 105.11 (2008), pp. 4387–4392. ISSN: 0027-8424.

[107] Margarita Salazar, Wanwipa Vongsangnak, Gianni Panagiotou, Mikael R. Andersen, and Jens Nielsen. "Uncovering transcriptional regulation of glycerol metabolism in Aspergilli through genome-wide gene expression data analysis". In: *Molecular Genetics and Genomics* 282.6 (2009), pp. 571–586. ISSN: 16174615.

[108] Wen-Bing Yin et al. "Discovery of Cryptic Polyketide Metabolites from Dermatophytes Using Heterologous Expression in Aspergillus nidulans". In: *ACS Chemical Biology* 2.11 (2013), pp. 629–634.

[109] Yi-Ming Chiang et al. "An Efficient System for Heterologous Expression of Secondary Metabolite Genes in Aspergillus nidulans". In: *Journal of the American Chemical Society* 135.20 (2013), pp. 7720–7731.

[110] Morten Thrane Nielsen et al. "Heterologous Reconstitution of the Intact Geodin Gene Cluster in Aspergillus nidulans through a Simple and Versatile PCR Based Approach". In: *PLoS ONE* 8.8 (2013), p. 72871. ISSN: 19326203.

[111] Peng Zhang et al. "A cryptic pigment biosynthetic pathway uncovered by heterologous expression is essential for conidial development in Pestalotiopsis fici". In: *Molecular Microbiology* 105.3 (Aug. 2017), pp. 469–483.

[112] Matthew T. Robey et al. "Identification of the First Diketomorpholine Biosynthetic Pathway Using FAC-MS Technology". In: *ACS Chemical Biology* 13.5 (2018), pp. 1142–1147. ISSN: 15548937.

[113] Jakob B. Nielsen, Michael L. Nielsen, and Uffe H. Mortensen. "Transient disruption of non-homologous end-joining facilitates targeted genome manipulations in the filamentous fungus Aspergillus nidulans". In: *Fungal Genetics and Biology* 45.3 (2008), pp. 165–170. ISSN: 10871845.

[114] Dorte Koefoed Holm. "Development and Implementation of Novel Genetic Tools for Investigation of FungalSecondary Metabolism". PhD thesis. Kgs. Lyngby: Technical University of Denmark, 2013.

[115] Jakob K. Rendsvig and Jakob B. Hoof. "Characterisation of bidirectional Aspergillus histone promoter (working title)". In: *Manuscript in preperation* ().

[116] Robert A. Samson, Jos Houbraken, Ulf Thrane, Jens C. Frisvad, and Birgitte Andersen. *Food and Indoor Fungi*. Utrecht, the Netherland : CBS-KNAW Fungal Biodiversity Centre, 2010, p. 390. ISBN: 978-90-70351-82-3.

[117] M. H H Nørholm. "A mutant Pfu DNA polymerase designed for advanced uracil-excision DNA engineering". In: *BMC Biotechnology* 10.21 (2010). ISSN: 14726750.

[118] Hussam H. Nour-Eldin, Fernando Geu-Flores, and Barbara A. Halkier. "Chapter 13: USER Cloning and USER Fusion: The Ideal Cloning Techniques for Small and Big Laboratories". In: *Plant Secondary Metabolism Engineering, Methods in Molecular Biology 643,* vol. 643. 2010, pp. 185–200. ISBN: 978-1-60761-722-8.

[119] Alexandros Stamatakis. “RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies”. In: *Bioinformatics* 30.9 (2014), pp. 1312–1313.

[120] Robert C Edgar. “MUSCLE: multiple sequence alignment with high accuracy and high throughput”. In: *Nucleic Acids Research* 32.5 (2004), pp. 1792–1797.

[121] J Castresana. “Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis”. In: *Mol. Biol. Evol* 17.4 (2000), pp. 540–552. ISSN: 0737-4038.

[122] Matthew D. Lebar et al. “Identification and functional analysis of the aspergillic acid gene cluster in Aspergillus flavus”. In: *Fungal Genetics and Biology* 116 (July 2018), pp. 14–23. ISSN: 10960937.

[123] Matthew J. Nicholson, Albert Koulman, Brendon J. Monahan, Beth L. Pritchard, Gary A. Payne, and Barry Scott. “Identification of two aflatrem biosynthesis gene loci in Aspergillus flavus and metabolic engineering of Penicillium paxilli to elucidate their function”. In: *Applied and Environmental Microbiology* 75.23 (2009), pp. 7469–7481. ISSN: 00992240.

[124] Rasmus Dam Wollenberg et al. “Chrysogine Biosynthesis Is Mediated by a Two-Module Nonribosomal Peptide Synthetase”. In: *Journal of Natural Products* 80.7 (2017), pp. 2131–2135. ISSN: 15206025.

[125] Choong Soo Yun, Takayuki Motoyama, and Hiroyuki Osada. “Biosynthesis of the mycotoxin tenuazonic acid by a fungal NRPS-PKS hybrid enzyme”. In: *Nature Communications* 6 (2015), p. 8758. ISSN: 20411723.

[126] Kenji Watanabe. *Effective use of heterologous hosts for characterization of biosynthetic enzymes allows production of natural products and promotes new natural product discovery*. Tech. rep. 12. 2014, pp. 1153–1165.

[127] Mitchell J Sullivan, Nicola K Petty, and Scott A Beatson. “Easyfig: a genome comparison visualizer”. In: *BIOINFORMATICS APPLICATIONS NOTE* 27.7 (2011), pp. 1009–101010.

A Supplementary section 3.1 – Paper I

Following is the supplementary material for the article I presented in Chapter 3 section 3.1 'Linking secondary metabolites to gene clusters through genome sequencing of six diverse *Aspergillus* species'.

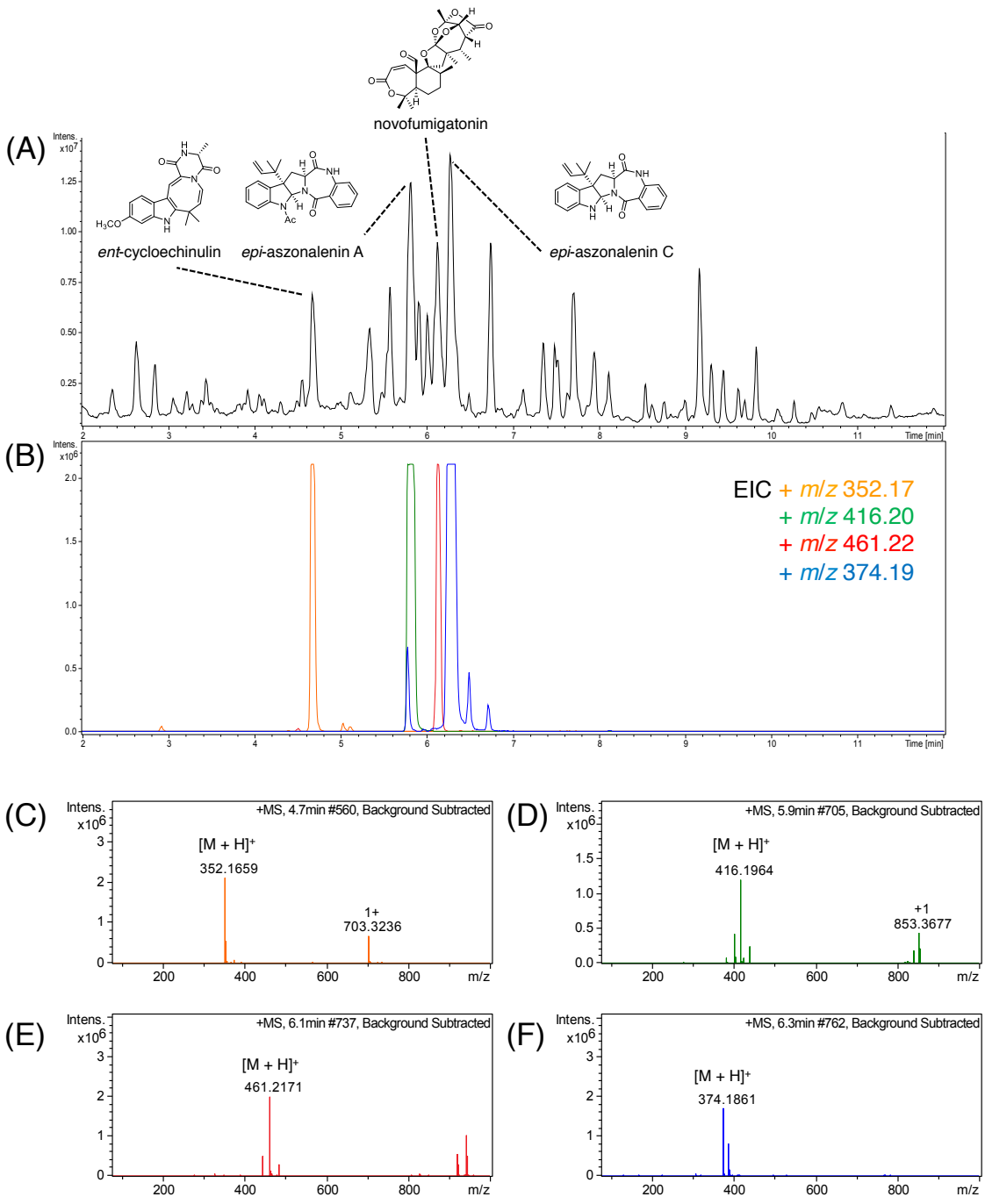


Figure S1 LC-MS analysis of the metabolites from Aspergillus novofumigatus IBT 16806 cultivated on YES agar medium. (A) Total and (B) extracted ion chromatograms, and mass spectra of (C) *ent*-cycloechinulin, (D) *epi*-aszonalenin A, (E) novofumigatonin, and (F) *epi*-aszonalenin A.

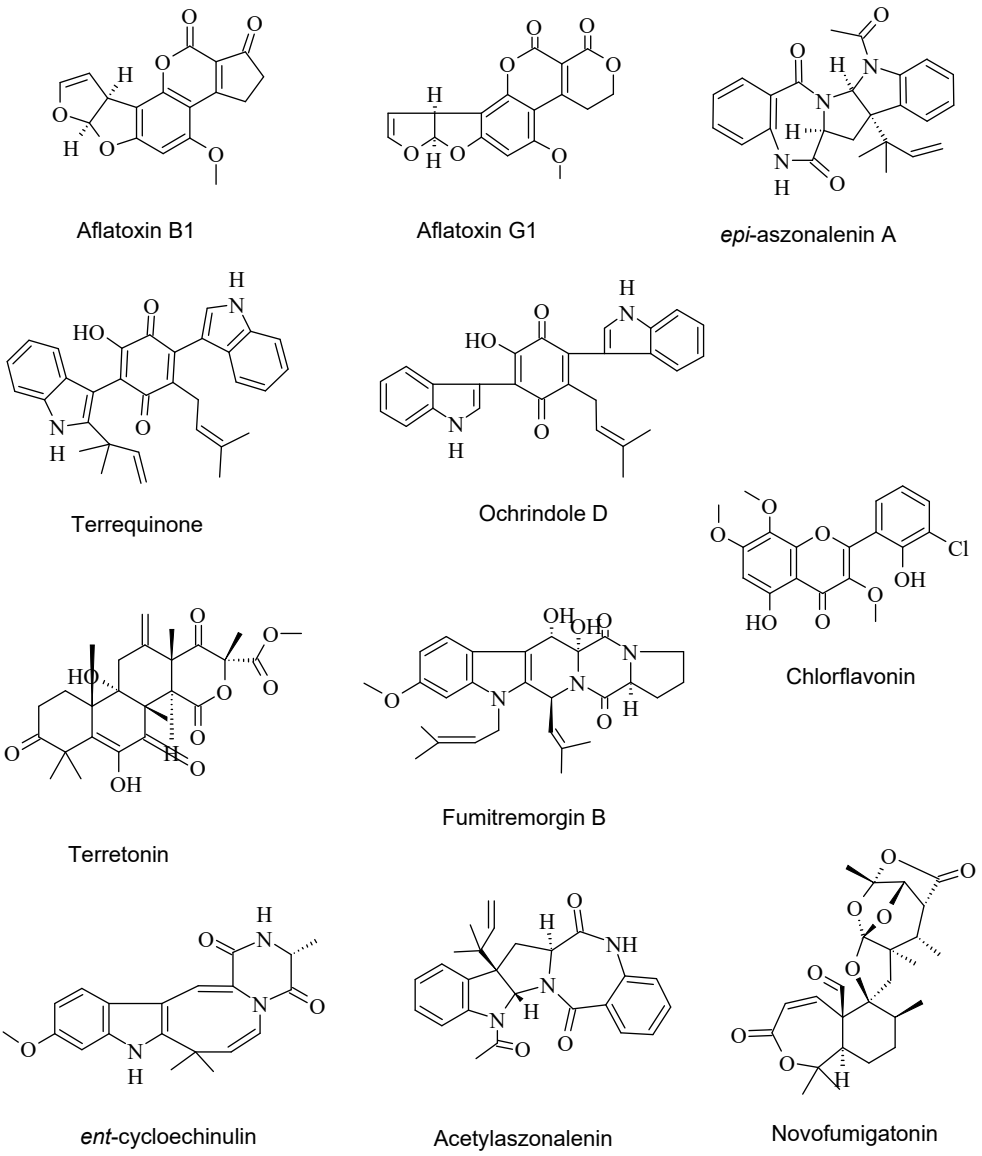


Figure S2 Overview of the chemical structures of compounds mentioned in the article (Aflatoxin B1 and G1, *epi*-aszonalenin A, terrequinone, ochrindole D, chlorflavonin, terretonin, fumitremorgin B, *ent*-cycloechinulin, acetylaszonlenin and novofumigatonin).

SI Text

Part 1: Materials and Methods

Strains:

A. campestris (CBS 348.81 = IBT28561, NCBI Taxonomy ID: 1392248) was first isolated in 1979 from soil collected in northern North Dakota. Initially is was placed in the *ochraceus* group (section *Circumdati*) [1]. This was however only based on morphology. In 2000 Rahbæk et al. suggested that *A. campestris* should be placed in section *Candidi* based on chemotaxonomical evidence [2]. In an investigation of section *Candidi* it was further consolidated that *A. campestris* belongs to this section based on among other phylogenetic studies of the calmodulin and β-tubulin genes [3]. *A. campestris* is known to produce a range of interesting chemical compounds including, candidusin C, terphenyllin and chlorflavonin [2, 3].

A. novofumigatus (CBS 117520 = IBT16806, NCBI Taxonomy ID: 1392255) was originally isolated Californian chamise chaparral soil collected in 1965. It is closely related to the pathogenic species *A. fumigatus* and belongs to the section *Fumigati*. It was initially suggested to be a separate species in 2005, this was mainly based on phylogenetic studies of the beta-tubulin, calmodulin and actin genes and on the extrolite profile [4]. *A. novofumigatus* has been recorded in one instance of aspergillosis along with another species in a patient with leukaemia [5].

A. ochraceoroseus (IBT 24754 = CBS 550.77, NCBI Taxonomy ID: 1392256) was first isolated from soil from the Taï national Park in the Ivory Coast in 1978 and based on the morphology it was assigned to ochraceus group (section *Circumdati*) [6]. Looking at the phylogeny of several housekeeping genes and the extrolite profile *A. ochraceoroseus* it is more closely related to subgenus *Nidulantes* and *Veriscolores* [7]. Currently it is believed to be a member of section Ochraceorosei [8, 9].

A. steynii (IBT 23096 = CBS 112812, NCBI Taxonomy ID: 1392250) has been isolated from Arabica green coffee bean from India. It belongs to section *Circumdati* and is closely related to *A. elegans*. *A. steynii* is important in food spoilage since it is known to produce ochratoxin A in addition to several other extrolites [10, 11].

A. candidus (CBS 102.13 = IBT 13984) was isolated in Japan and belongs to section *Candidi* [2, 3].

A. taichungensis (NCBI 482145 =IBT19404, NCBI Taxonomy ID: 482145) was isolated from Soil in Taiwan and belongs to section *Candidi* [2, 3].

The strains used in this study can be seen in Table 1.

Table 1 The species with whole genome sequences used in this study.

Species	Collection number	JGI abbreviation	Reference
A. campestris	CBS 348.81 = IBT28561	Aspcam1	This study
A. novofumigatus	CBS 117520 = IBT16806	Aspnov1	This study
A. ochraceoroseus	CBS 550.77 = IBT 24754	Aspoch1	This study
A. steynii	CBS 112812 = IBT 23096	Aspste1	This study
A. candidus	CBS 102.13 = IBT 13984	Aspcan1	This study
A. taichungensis	NCBI 482145 =IBT19404	Asptaic1	This study
A. oryzae	RIB40	Aspor1	[12, 13]
A. flavus	NRRL3357	Aspfl1	[12]
A. nidulans	FGSC A4	Aspnid1	[12, 14]
A. niger	ATCC 1015	Aspni7	[15]
A. fumigatus Af293	FGSC A1100	Aspfu1	[16]
A. fumigatus A1163	FGSC A1163	Aspfu_A1163 _1	[16–19]
N. fischeri/A. fischerianus	NRRL 181	Neofi1	[12, 20]
A. clavatus	NRRL 1	Aspcl1	[12]
A. terreus	NIH2624	Aspte1	[12]
N. crassa	OR74A	Neucr2	[21]
P. chrysogenum		Pench1	These sequence data were produced by the US Department of Energy Joint Genome Institute

			http://www.jgi.doe.gov/ in collaboration with the user community.

Growth of strains and DNA extraction

Biomass for all fungal strains was obtained from shake flasks containing 200ml of complex media CYA [22]. Biomass was isolated by filtering through Miracloth (Millipore, 475855-1R), freeze dried, and stored at -80C. Subsequently, a sample of frozen biomass was used for RNA purification. First hyphae were lysed in a 2ml micro tube, together with a 5mm Steel bead (QIAGEN), placed in liquid nitrogen by using the QIAGEN TissueLyser LT at 45 Hz for 50 seconds. Then the QIAGEN RNeasy mini Plus Kit was used, and the RLT Plus buffer (with 2-mercaptoethanol) was added to the samples, vortexed and spun down. The lysate was then used in step 4 in the instructions provided by the manufacturer, and the protocol was followed from this step. For genomic DNA a protocol inspired by Fulton et al. [23] was used. For details see Additional file 8.

Genome sequencing and assembly. *A. campestris, A. novofumigatus, A. ochraceoroseus* and *A. steynii* were whole-genome sequenced using PacBio RS and assembled using Hierarchical-Based AssembleR (HBAR) which is a developing version derived from Hierarchical Genome Assembly Process (HGAP) [24] allowing larger genomes to be processed.

Unamplified libraries were generated using Pacific Biosciences standard template preparation protocol for creating >10kb libraries. 5ug of gDNA was used to generate each library and the DNA was sheared using Covaris g-Tubes to generate sheared fragments of >10kb in length. A modified version of the protocol was used for 5kb PacBio libraries using a Covaris LE220 focused-ultrasonicator with their Red miniTUBES for DNA shearing. The sheared DNA fragments were then prepared using Pacific Biosciences SMRTbell template preparation kit, where the fragments were treated with DNA damage repair, had their ends repaired so that they were blunt-ended, and 5' phosphorylated. Pacific Biosciences hairpin adapters were then ligated to the

fragments to create the SMRTbell template for sequencing. The SMRTbell templates were then purified using exonuclease treatments and size-selected using AMPure PB beads. Sequencing primer was then annealed to the SMRTbell templates and Version P4 sequencing polymerase was bound to them. The prepared SMRTbell template libraries were then sequenced on a Pacific Biosciences RSII sequencer using Version C2 chemistry and 2 hour sequencing movie run times. Genomes of *A. campestris, A. novofumigatus, A. ochraceoroseus* and *A. steynii* were assembled using HBAR (https://github.com/PacificBiosciences/HBAR-DTK) and polished with quiver.

A. candidus and *A. taichungensis* were whole-genome sequenced using Illumina. 100ng of DNA was sheared to 270bp using the Covaris LE220 (Covaris) and size selected using SPRI beads (Beckman Coulter). The fragments were treated with end-repair, A-tailing, and ligation of Illumina compatible adapters (IDT, Inc) using the KAPA-Illumina library creation kit (KAPA biosystems). qPCR was used to determine the concentration of the libraries. Libraries were sequenced on the Illumina Hiseq in 2x150bp format.
The assemblies were produced using combination of Velvet and AllPathsLG v.R47710 assemblers. The raw fastq file was QC filtered to remove artifact/process contamination and then separated into two fastq files, one with mitochondrial data only and the remainder in the target fastq. The target fastq was subsequently assembled using Velvet [25]. The resulting assembly was used to create a long mate-pair library with insert 3000 +/- 90 bp which was then assembled together with the target fastq using AllPathsLG release version R47710, [26].

Transcriptome sequencing and assembly

Stranded cDNA libraries were generated using the Illumina Truseq Stranded RNA LT kits. mRNA was purified from 1µg (100ng for WPOO) of total RNA using magnetic beads containing poly-T oligos. mRNA was fragmented using divalent cations and high temperature. The fragmented RNA was reversed transcribed

using random hexamers and SSII (Invitrogen) followed by second strand synthesis. The fragmented cDNA was treated with end-pair, A-tailing, adapter ligation, and 10 (12 for WPOO) cycles of PCR. qPCR was used to determine the concentration of the libraries. Libraries were sequenced on the Illumina Hiseq. Illumina reads of stranded RNA-seq data were used as input for de novo assembly of RNA contigs. Reads were assembled into consensus sequences using Rnnotator (v. 3.3.1 or later), which consists of three major components: preprocessing of reads, assembly, and postprocessing of contigs [27].

Genome annotation. Annotation of the genomes was completed using the JGI annotation pipeline and made publicly available via JGI fungal genome portal Mycocosm [28, 29].

Also the software tool Secondary Metabolite Unknown Regions Finder (SMURF) was used to predict secondary metabolite gene clusters (SMGC) in the genomes [30].

Sequences were analysed using InterProScan5 in order to investigate potential protein functions [31].

BLASTP. Each protein in each of the genomes has been compared to all other proteins using the BLASTP function from the BLAST+suite version 2.2.27 with a non-restrictive e-value cutoff of 10^{10} [32, 33].

Phylogenetic analysis – CVTree. A phylogenetic tree was constructed using the Composition Vector approach. The web based server CVTree3 was used (http://tlife.fudan.edu.cn/archaea/cvtree/cvtree3/) [34, 35]. The proteome for each of the species were uploaded in fasta format and the K-tuple length was set to 8 and then the project as run resulting in a phylogenetic tree. The tree was visualized using FigTree v1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/).

Identification of species-specific genes

All predicted sets of protein sequences for the 4 genomes in this paper (A. campestris, A. novofumigtus, A.

ochraceoroseus, A. steynii) and the nine reference genomes (*A. nidulans, A. oryzae, A. flavus, A. niger, A. fumigatus Af293, A. fumigatus A1163, N. fischeri, A. clavatus, A.terreus*), and two outgroup species (*N. crassa and P. chrysogenum*) were aligned using the BLASTp function from the BLAST+ suite version 2.2.27 with an e-value <= 10^-10 [32, 33]. These 225 whole-genome BLAST tables were analyzed to identify bi-directional hits in all pairwise comparisons. Using custom Python-scripts, paralogs were first identified within the same genome and grouped into sequence similar families using single linkage, meeting the criterion; the sum of the alignment coverage between the pairwise sequences ≥ 130%, the alignment identity between the pairwise sequences ≥ 50%, and the pairwise hit must be bi-directional (present in both BLAST directions). The orthologs were identified across genomes and grouped into sequence similar families using single linkage meeting the same criterions. Singletons were assigned a family having only one gene member. This allowed for identification of species unique genes. All homologs were assigned functional and structural domains using InterPro version 48 [36, 37] and checked for annotation and sequencing errors by investigating scaffold location and sequence identity.

N6-methyldeoxyadenine analysis

Modification detection was performed using single-molecule real-time (SMRT) sequencing and the PacBio SMRT Analysis 2.3.0 toolkit [38]. Following modification detection, 6mA sites were filtered following Mondo et al., 2017, which includes filtering modifications by both coverage (minimum 15x, maximum determined using the R-boxplot function) and modification quality value (mQV; minimum mQV = 25).

Synteny plots of SMGC – Easyfig. Synteny plots were made using Easyfig version 2.1 (http://easyfig.sourceforge.net) [39]. All GenBank files describing the clusters to be compared were uploaded. The tBLASTx option was used with an e-≤ 0.001 , an alignment length ≥ 50 and an alignment identity ≥ 35. Length of scale was set to 2000 bp under the figure option. If necessary, the sequence can be

reversed under the sub region options. After the options have been set the figure can be created choosing vector format (svg).

Homology of *A. novofumigatus* and *A. fumigatus*

The number of proteins from *A. novofumigatus* with homologs in *A. fumigatus* was determined based on BLASP hit with identity ≥50% and the sum of the query and hit coverage ≥130%.

Synteny of *A. novofumigatus* and *A. fumigatus*

The syneteny between *A. novofumigatus* and *A. fumigatus* was determined using MUMmer (http://mummer.sourceforge.net) [40–42]. 'NUCmer' (NUCleotide MUMmer) was used with standard setting to generate alignments between the two species, followed by 'show-coords' to generate a file with the output from where the total coverage of the genome, maximum block length and mean block length could be calculated based on the *A. novofumigtus* length of the alignments.

Comparison of *A. fumigatus* and *A. novofumigatus* secondary metabolite gene clusters

Comparison of the best hits for *A. fumigatus* and *A. novofumigatus* secondary metabolite gene clusters in other species, based on average percent identity of the backbone proteins. First, BLASTP comparisons of all *A. fumigatus* and *A. novofumigatus* SMGC backbone proteins against all SMGC backbone proteins form the dataset were created. Subsequently, an average of backbone proteins identity was calculated per cluster, the best hit is shown in the heatmap. Backbone proteins are defined as proteins with the annotations PKS(-like), NRPS(-like), hybrid, DMATS and TC.

LC-MS analysis of metabolites from *A. novofumigatus* IBT 16806

Aspergillus novofumigatus IBT 16806 was cultivated on YES agar plate at 25 °C for 7 days, and extracted with ethyl acetate containing 1% formic acid. The extract for LC-MS analysis were injected into a Dionex Ultimate 3000 UHPLC system (Thermo Scientific) - a maXis 3G QTOF orthogonal mass spectrometer (Bruker Daltonics), using Electrospray Ionization with a Kinetex C_{18} column (2.1 i.d. x 100 mm; Phenomenex). Separation was performed with a solvent system of water containing 20 mM formic acid (solvent A) and acetonitrile containing 20 mM formic acid (solvent B), at a flow rate of 0.4 ml/min and a column temperature of 40 °C, using the following program: a linear gradient from 10:90 (solvent B/solvent A) to 100:0 for 10 min, 100:0 for the following 3 min, and a linear gradient from 100:0 to 10:90 within the following 2 min. To illustrate our approach of linking metabolites produced by *A. novofumigatus* to their respective gene clusters, we chose to target our metabolite analysis towards the model compounds novofumigatonin, *ent*-cycloechinulin, *epi*-aszonalenin A and C, since they represent major metabolites produced by this species and because we have them as pure standards in our in-house collection of fungal metabolites [43]. For a full list of metabolites known from *A. novofumigatus* please consult Frisvad & Larsen [44].

References

1. Christensen M. The Aspergillus ochraceus Group: Two New Species from Western Soils and a Synoptic Key. Mycologia. 1982;74:210–25. http://www.jstor.org.proxy.findit.dtu.dk/stable/3792887?origin=crossref&seq=1#page_scan_tab_contents. Accessed 20 Oct 2015.

2. Rahbæk L, Frisvad JC, Christophersen C. An amendment of Aspergillus section Candidi based on chemotaxonomical evidence. Phytochemistry. 2000;53:581–6. doi:10.1016/S0031-9422(99)00596-8.

3. Varga J, Frisvad JC, Samson RA. Polyphasic taxonomy of Aspergillus section Candidi based on molecular, morphological and physiological data. Stud Mycol. 2007;59:75–88.

doi:10.3114/sim.2007.59.10.

4. Hong S-B, Go S-J, Shin H-D, Frisvad JC, Samson RA. Polyphasic Taxonomy of Aspergillus fumigatus and Related Species. Mycologia. 2005;97:1316–29. http://www.jstor.org.proxy.findit.dtu.dk/stable/3762370?seq=1#page_scan_tab_contents. Accessed 20 Oct 2015.

5. Peláez T, Álvarez-Pérez S, Mellado E, Serrano D, Valerio M, Blanco JL, et al. Invasive aspergillosis caused by cryptic Aspergillus species: A report of two consecutive episodes in a patient with leukaemia. J Med Microbiol. 2013;62 PART3:474–8.

6. Bartoli A, Maggi O. Four new species of Aspergillus from Ivory Coast soil. Trans Br Mycol Soc. 1978;71:383–94. doi:10.1016/S0007-1536(78)80064-3.

7. Klich M a, Cary JW, Beltz SB, Bennett C a. Phylogenetic and morphological analysis of Aspergillus ochraceoroseus. Mycologia. 2003;95:1252–60.

8. Cary JW, Harris-Coward PY, Ehrlich KC, Moore GG, Wei Q, Bhatnagar D. Functional and phylogenetic analysis of the Aspergillus ochraceoroseus aflQ (ordA) gene ortholog. Mycologia. 2012;104:857–64.

9. Samson RA, Visagie CM, Houbraken J, Hong S-B, Hubka V, Klaassen CHW, et al. Phylogeny,identification and nomenclature of the genus Aspergillus. Stud Mycol. 2014;78:141–73. doi:10.1016/j.simyco.2014.09.001.

10. Visagie CM, Houbraken J, Frisvad JC, Hong S, Klaassen CHW, Perrone G, et al. Ochratoxin production and taxonomy of the yellow aspergilli (Aspergillus section Circumdati). Stud Mycol. 2014;78:1–61. doi:10.1016/j.simyco.2014.09.001.

11. Frisvad JC, Frank J., Houbraken JAMP, Kuijpers AFA, Samson RA, Frisvad JC. New ochratoxin A

producing species of Aspergillus section Circumdati. Stud Mycol. 2004;50:23–43.

12. Arnaud MB, Cerqueira GC, Inglis DO, Skrzypek MS, Binkley J, Chibucos MC, et al. The Aspergillus Genome Database (AspGD): Recent developments in comprehensive multispecies curation, comparative genomics and community resources. Nucleic Acids Res. 2012;40:653–9.

13. Machida M, Asai K, Sano M, Tanaka T, Kumagai T, Terai G, et al. Genome sequencing and analysis of Aspergillus oryzae. Nature. 2005;438:1157–61. doi:10.1038/nature04300.

14. Galagan JE, Calvo SE, Cuomo C, Ma L-J, Wortman JR, Batzoglou S, et al. Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae. Nature. 2005;438:1105–15. doi:10.1038/nature04341.

15. Andersen MR, Salazar MP, Schaap PJ, Van De Vondervoort PJI, Culley D, Thykaer J, et al. Comparative genomics of citric-acid-producing Aspergillus niger ATCC 1015 versus enzyme-producing CBS 513.88. Genome Res. 2011;21:885–97.

16. Nierman WC, Pain A, Anderson MJ, Wortman JR, Kim HS, Arroyo J, et al. Genomic sequence of the pathogenic and allergenic filamentous fungus Aspergillus fumigatus. Nature. 2005;438:1151–6. doi:10.1038/nature04332.

17. Fedorova ND, Khaldi N, Joardar VS, Maiti R, Amedeo P, Anderson MJ, et al. Genomic islands in the pathogenic filamentous fungus Aspergillus fumigatus. PLoS Genet. 2008;4:e1000046. doi:10.1371/journal.pgen.1000046.

18. Ronning CM, Fedorova ND, Bowyer P, Coulson R, Goldman G, Kim HS, et al. Genomics of Aspergillus fumigatus. Rev Iberoam Micol. 2005;22:223–8. http://www.ncbi.nlm.nih.gov/pubmed/16499415. Accessed 10 Dec 2015.

19. Joardar V, Abrams NF, Hostetler J, Paukstelis PJ, Pakala S, Pakala SB, et al. Sequencing of

mitochondrial genomes of nine Aspergillus and Penicillium species identifies mobile introns and accessory genes as main sources of genome size variability. BMC Genomics. 2012;13:698. doi:10.1186/1471-2164-13-698.

20. Lonial S, Williams L, Carrum G, Ostrowski M, McCarthy P. Neosartorya fischeri: an invasive fungal pathogen in an allogeneic bone marrow transplant patient. Bone Marrow Transplant. 1997;19:753–5. doi:10.1038/sj.bmt.1700715.

21. Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, et al. The genome sequence of the filamentous fungus Neurospora crassa. Nature. 2003;422:859–68. doi:10.1038/nature01554.

22. Samson RA, Houbraken J, Thrane U, Frisvad JC, Andersen B. Food and Indoor Fungi. CBS-KNAW Fungal Biodiversity Centre; 2010. http://findit.dtu.dk/en/catalog/2185758085. Accessed 14 Aug 2017.

23. Fulton TM, Chunwongse J, Tanksley SD. Microprep protocol for extraction of DNA from tomato and other herbaceous plants. Plant Mol Biol Report. 1995;13:207–9.

24. Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013;10:563–9. doi:10.1038/nmeth.2474.

25. Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18:821–9. doi:10.1101/gr.074492.107.

26. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A. 2011;108:1513–8. doi:10.1073/pnas.1017351108.

27. Martin J, Bruno VM, Fang Z, Meng X, Blow M, Zhang T, et al. Rnnotator : an automated de

novo transcriptome assembly pipeline from stranded RNA-Seq reads Duplicate read removal Multiple Velvet assemblies. 2010.

28. Grigoriev I V., Martinez DA, Salamov AA. Fungal genomic annotation. Appl Mycol Biotechnol. 2006;6 C:123–42.

29. Grigoriev I V., Nikitin R, Haridas S, Kuo A, Ohm R, Otillar R, et al. MycoCosm portal: Gearing up for 1000 fungal genomes. Nucleic Acids Res. 2014;42:699–704.

30. Khaldi N, Seifuddin FT, Turner G, Haft D, Nierman WC, Wolfe KH, et al. SMURF: Genomic mapping of fungal secondary metabolite clusters. Fungal Genet Biol. 2010;47:736–41. doi:10.1016/j.fgb.2010.06.003.

31. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30:1236–40. doi:10.1093/bioinformatics/btu031.

32. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10. doi:10.1016/S0022-2836(05)80360-2.

33. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421. doi:10.1186/1471-2105-10-421.

34. Qi J, Wang B, Hao BI. Whole Proteome Prokaryote Phylogeny Without Sequence Alignment: A K-String Composition Approach. J Mol Evol. 2004;58:1–11.

35. Zuo G, Li Q, Hao B. On K-peptide length in composition vector phylogeny of prokaryotes. Comput Biol Chem. 2014;53:166–73. doi:10.1016/j.compbiolchem.2014.08.021.

36. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, et al. InterPro: The integrative protein signature database. Nucleic Acids Res. 2009;37 SUPPL. 1:211–5.

37. Mitchell A, Chang H-Y, Daugherty L, Fraser M, Hunter S, Lopez R, et al. The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Res. 2014;43:D213-221. doi:10.1093/nar/gku1243.

38. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods. 2010;7:461–5. doi:10.1038/nmeth.1459.

39. Sullivan MJ, Petty NK, Beatson SA. Easyfig: a genome comparison visualizer. Bioinformatics. 2011;27:1009–10. doi:10.1093/bioinformatics/btr039.

40. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004;5:R12.

41. Delcher AL, Phillippy A, Carlton J, Salzberg SL. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res. 2002;30:2478–83. doi:10.1093/nar/30.11.2478.

42. Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, Salzberg SL. Alignment of whole genomes. Nucleic Acids Res. 1999;27. http://mummer.sourceforge.net/MUMmer.pdf. Accessed 24 Nov 2017.

43. Nielsen KF, Månsson M, Rank C, Frisvad JC, Larsen TO. Dereplication of Microbial Natural Products by LC-DAD-TOFMS. J Nat Prod. 2011;74:2338–48. doi:10.1021/np200254t.

44. Frisvad JC, Larsen TO. Extrolites of Aspergillus fumigatus and Other Pathogenic Species in Aspergillus Section Fumigati. Front Microbiol. 2016;6 January:1–14.

Part 2: Protocol for preparation of Fungal DNA

The protocol described below has successfully been employed to isolate genome-sequencing grade genomic DNA for more than 200 different *Aspergillus* species.

List of Materials:

D-Sorbitol (Sigma, S1876 – CAS 50-70-4)
Tris-Base (Sigma 7-9, T1378 – CAS 7786-1)
37% HCl (Th. Geyer, 836,1000)
EDTA (Merck, 324503 – CAS 6381-92-6)
Sodium Cloride (NaCl) (AppliChem A1371,9010 – CAS 7647-14-5]
Cetyl trimethylammonium bromide (CTAB) (Sigma 52365 – CAS 57-09-0)
Sarkosyl NL (Sigma, L5777 – CAS 137-16-6)
Polyvinylpyrrolidone (PVP) (Sigma PVP-40T – CAS 9003-39-8)
Proteinase K (NEB P8107S)
Potassium acetate (J. T. Baker 0129910025 – CAS 127-08-2)
Phenol:Chloroform:Isoamylalcohol (25:24:1) (Sigma P3803)
Sodium acetate (J. T. Baker 9914011001 – CAS 6131-90-4]
96 % Ethanol (vwr chemicals)
70 % Ethanol (vwr chemicals)
Isopropanol (Merck, 109634 – CAS 67-63-0)
Liquid nitrogen
Sodium Hydroxide (Sigma S5881 – CAS 1310-73-2)
RNase A (Sigma R-4875 – CAS 9001-99-4)

Preparation of liquid Media

Buffers:

- 5M Potassium acetate (pH 7.5): 122.5 g potassium acetate and ddH_2O up to 250 mL. pH adjusted with acetic acid.
- 3M Sodium acetate: 81.65 g sodium acetate and ddH_2O up to 200 mL.
- 1% PVP: 2 g PVP in 200 mL ddH_2O
- 5% Sarkosyl: 10 g Sarkosyl in 200 mL ddH_2O.
- 1M Tris-HCl (pH 9): 60.57 g Tris-base and 4.81 ml 37% HCl. Add ddH_2O up to 500 mL.
- 0.5M EDTA: 116.4 g EDTA. Add ddH_2O up to 500 mL. Add Sodium Hydroxide pellet until pH reach 8.0.
- Buffer A: 31.9 g Sorbitol, 50 mL 1M Tris-HCl (pH 9), 5 mL 0.5M EDTA (pH 8) and ddH_2O up to 500 mL.
- Buffer B: 100 mL 1M Tris-HCl (pH 9), 50 mL 0.5M EDTA, 58.44 g NaCl, 10 g CTAB. Add ddH_2O up to 500 mL.
- TE (pH 9): 1.21 g Tris-base, 0.37 g EDTA. Add ddH_2O up to 1000 mL.

All solutions above are to be autoclaved!

- Lysis Buffer: For 10 ml pr. sample use: 3.75 ml Buffer A; 3.75 ml Buffer B; 1.5 ml 5 % Sarkosyl; 1 ml 1 % PVP; 100 µl Proteinase K
- RNase A: Dissolve 10 mg dry powder in 1 ml ddH_2O

Equipment:

- Nanodrop: NanoDrop ND 1000 Spectrophotometer or NanoDrop Lite from Qiagen.
- Qubit: Qubit 1.0 fluorometer from Invitrogen and Qubit dsDNA BR Assay Kit (Q32853) from ThermoFisher.
- Mortar and pestle.
- Centrifuge for 50ml Falkon tubes at 4°C.

Protocol:

1. Pre-heat Buffer B at 65°C
2. Prepare Lysis Buffer just before use and keep at 65°C.
3. Transfer freeze-dried mycelia into a mortar and cover with liquid nitrogen. Grind material and transfer to a 50 ml Falcon tube as soon as all liquid nitrogen has evaporated. Powder in the tube should not exceed the 5ml mark, but a minimum of 3 ml is recommended. Note powder must not thaw.
4. Add 10 ml Lysis Buffer and mix vigorously by vortexing.
5. Incubate for 30 min at 65°C. Mix frequently by inverting the tube.
6. Add 3.35 ml 5 M Potassium acetate. Mix gently by inverting the tube 5-7 times. Incubate solution 30 min on ice.
7. Centrifuge for 30 min at 5,000 g at 4°C.
8. Transfer the supernatant, approximately 9mL, to a new 50 ml Falcon tube and add 5ml of Phenol:Chloroform:Isoamylalcohol (25:24:1). Mix gently 5-7 times.
9. Centrifuge 20 min at 4,000 g at 4 °C.
10. Transfer the aqueous phase (~8mL) to a new 50 ml Falcon tube. Note avoid any material from the interphase.
11. Add 100 µl RNase A (10 mg/ml) and mix gently. Incubate at room temperature for 30-60 min.
12. Add 1/10 volume of 3M Sodium acetate and 1 volume ice-cold 96% Ethanol (Alternatively, Isopropanol can be used, but it may adversely influence A260/A280 measurements). Incubate solution at 20°C for 30 min.
13. Centrifuge for 30 min at 10,000 g and 4°C
14. Discard the supernatant.
15. Wash the pellet with 2 ml 70 % ethanol and pipette as much away without disturbing the pellet.

16. Dry the pellet at room temperature until all ethanol has evaporated (approximately 15 minutes).
    Note: do not let the pellet dry out!
17. Dissolve the pellet in 500 µL TE. This may take over-night incubation at room temperature with light shaking. Transfer DNA solution to a 2 ml Eppendorf tube.
18. Take a sample for DNA quality assessments (see below) and store the remaining DNA solution at -20 °C until further use.
19. For testing DNA quality:
    Make a 20-fold dilution of the DNA solution (from step 18) in a 1.5 ml Eppendorf tube to a total volume of 100 µl.
    A. Run a 5-10 µl diluted sample on an agarose gel to estimate the quality and concentration.
    B. Use the nanodrop for A_{260}/A_{280} measurements. Ratios should be in the range of 1.6-2.2.
    C. Use the Qubit to determine DNA concentration estimations. Good predations fall in the range of 20-200 ng/µl DNA in stock solution.

Table S1 Overview of the most common InterPro domains for the unique genes in the investigated species and the number of times the InterPro domain is found in a unique gene in the investigated species.

Species	Total predicted proteins	Total unique proteins	Unique genes %	InterPro annotated unique	Most common IPR, # of IPR in unique genes						
					IPR001138	IPR016040	IPR002110	IPR016196	IPR011009	IPR007219	IPR001128
A. campestris	9764	2162	22%	670 (31%)	43	34	35	21	19	10	28
A. clavatus	9121	1053	12%	349 (33%)	20	23	18	20	6	7	21
A. flavus	12604	1953	15%	659 (34%)	52	46	32	48	25	37	22
A. fumigatus Af293	9781	188	2%	67 (36%)	0	2	5	3	1	0	1
A. fumigatus A1163	9916	343	3%	100 (29%)	8	1	8	5	7	3	3
A. niger ATCC 1015	11910	3168	27%	1232 (39%)	135	94	63	71	37	75	47
A. nidulans	10680	2391	22%	943 (39%)	103	63	37	46	17	47	36
A. novofumigatus	11549	1695	15%	462 (27%)	26	37	30	20	14	13	36
A. ochraceoroseus	8924	1881	21%	519 (28%)	57	26	6	12	27	20	18
A. oryzae	12031	1842	15%	635 (34%)	36	26	46	32	22	28	16
A. steynii	13211	3520	27%	1270 (36%)	163	114	67	54	27	72	43
A. terreus	10406	2117	20%	905 (43%)	55	66	54	45	37	52	37
A. fischerianus	10406	704	7%	249 (35%)	6	21	23	14	11	6	9
N. Crassa	10785	7063	65%	3471 (49%)	147	91	53	66	81	68	28
P. Chrysogenum		3215		1101 (34%)	126	64	44	35	57	52	30

IPR ID	Description
IPR001138	Fungal transcriptional regulatory protein, N-terminal
IPR016040	NAD(P)-binding
IPR002110	Ankyrin
IPR016196	MFS general substrate transporter
IPR011009	Protein kinase-like
IPR007219	Fungal specific transcription factor
IPR001128	Cytochrome P450

Top	IPR with most counts
Second	IPR with second most counts
Third	IPR with third most counts

Table S2 Allergens from *A. novofumigatus.* A list of allergens from *A. fumigatus* (from www.allergome.org) and the orthologs in *A. novofumigatus* including the percent identity of the BLAST comparison.

Allergen name	*A. fumigatus* AF293 accession	*A. novofumigatus* orthologue	% ID
Asp_f1	AFUA_5G02330	jgi-Aspnov1-365359-e_gw1.2.4275.1	98.86
Asp_f2	AFUA_4G09580	jgi-Aspnov1-432190-fgenesh1_pg.5_#_493	93.56
Asp_f3	AFUA_6G02280	jgi-Aspnov1-388041-estExt_Genewise1Plus.C_3_t50474	97.62
Asp_f4	AFUA_2G03830	jgi-Aspnov1-432819-fgenesh1_pg.6_#_39	94
Asp_f5	AFUA_8G07080	jgi-Aspnov1-395206-estExt_Genewise1Plus.C_7_t30340	95.58
Asp_f6	AFUA_1G14550	jgi-Aspnov1-512129-estExt_fgenesh1_pm.C_4_t10311	99.47
Asp_f7	AFUA_4G06670	jgi-Aspnov1-443000-fgenesh1_kg.5_#_844_#_Locus964v1rpkm138.38	92.6
Asp_f8	AFUA_2G10100	jgi-Aspnov1-30553-CE30552_16949	90
Asp_f9	AFUA_1G16190	jgi-Aspnov1-430704-fgenesh1_pg.4_#_185	94.07
Asp_f10	AFUA_5G13300	jgi-Aspnov1-363785-e_gw1.2.2138.1	95.19
Asp_f11	AFUA_2G03720	jgi-Aspnov1-426603-fgenesh1_pg.1_#_378	93.57
Asp_f12	AFUA_5G04170	jgi-Aspnov1-509883-estExt_fgenesh1_pm.C_2_t10392	99.27
Asp_f13	AFUA_4G11800	jgi-Aspnov1-431992-fgenesh1_pg.5_#_295	94.29
Asp_f15	AFUA_2G12630	jgi-Aspnov1-430704-fgenesh1_pg.4_#_185	94.07
Asp_f17	AFUA_4G03240	jgi-Aspnov1-409635-estExt_Genewise1.C_5_t50054	91.33
Asp_f18	AFUA_4G11800	jgi-Aspnov1-431992-fgenesh1_pg.5_#_295	94.3
Asp_f22	AFUA_6G06770	jgi-Aspnov1-367285-e_gw1.3.3588.1	99.09
Asp_f23	AFUA_2G11850	jgi-Aspnov1-447552-estExt_Genemark1.C_1_t30040	94.23
Asp_f26	AFUA_1G06830	jgi-Aspnov1-370658-e_gw1.4.1124.1	86.49
Asp_f27	AFUA_3G07430	jgi-Aspnov1-453038-estExt_Genemark1.C_6_t10423	90.8
Asp_f28	AFUA_6G10300	jgi-Aspnov1-367561-e_gw1.3.3658.1	96.1
Asp_f29	AFUA_5G11320	jgi-Aspnov1-445073-fgenesh1_kg.8_#_93_#_Locus606v1rpkm225.43	74.26
Asp_AfCalAp	AFUA_3G09690	jgi-Aspnov1-515921-estExt_fgenesh1_pm.C_100047	94.79
Asp_f_chitosanase	AFUA_4G01290	jgi-Aspnov1-412759-estExt_Genewise1.C_8_t10196	94.21
Asp_f_AT	AFUA_1G09470	jgi-Aspnov1-512574-estExt_fgenesh1_pm.C_4_t20289	91.19
Asp_f_catalase	AFUA_3G02270	jgi-Aspnov1-440530-fgenesh1_kg.3_#_979_#_Locus6690v1rpkm8.88	50.21
Asp_f_DPPV	AFUA_2G09030	jgi-Aspnov1-453870-estExt_Genemark1.C_7_t10446	88.57
Asp_f_glucosidase	AFUA_1G05770	jgi-Aspnov1-431600-fgenesh1_pg.4_#_1081	93.36
Asp_f_GST	AFUA_6G09690	jgi-Aspnov1-458931-fgenesh1_pm.3_#_489	96.67
Asp_f_GT	AFUA_6G11390	jgi-Aspnov1-398517-estExt_Genewise1.C_1_t10227	88.44
Asp_f_IAO	AFUA_6G03620	jgi-Aspnov1-413975-estExt_Genewise1.C_9_t20169	51.01
Asp_f_IPMI	AFUA_2G11260	jgi-Aspnov1-500519-estExt_fgenesh1_pg.C_1_t20448	96.27
Asp_f_LPL1	AFUA_4G08720	jgi-Aspnov1-516331-estExt_fgenesh1_pm.C_120063	94.96
Asp_f_LPL3	AFUA_3G14680	jgi-Aspnov1-516331-estExt_fgenesh1_pm.C_120063	90.83
Asp_f_mannosidase	AFUA_1G14560	jgi-Aspnov1-460012-fgenesh1_pm.4_#_314	94.65
Asp_f_MDH	AFUA_7G05740	jgi-Aspnov1-462836-fgenesh1_pm.7_#_121	96.19
Asp_f_PL	AFUA_2G00760	jgi-Aspnov1-499715-estExt_fgenesh1_pg.C_1_t10092	92.83
Asp_f_PUP	AFUA_5G03520	jgi-Aspnov1-438430-fgenesh1_kg.2_#_311_#_Locus9805v1rpkm2.72	95.05
Asp_f_SXR	AFUA_2G15430	jgi-Aspnov1-47013-CE47012_27880	99.57
Asp_f_CP	AFUA_8G01670	jgi-Aspnov1-445108-fgenesh1_kg.8_#_128_#_Locus912v1rpkm147.61	96.31
Asp_f_FDH	AFUA_6G04920	jgi-Aspnov1-440417-fgenesh1_kg.3_#_866_#_Locus1540v1rpkm84.43	95.16

Table S3 Virulence factors from *A. fumigatus*. A collection of virulence factors known from *A. fumigatus* and the best orthologs *in A. novofumigatus* along with the percent identity of the BLAST comparison.

***A. fumigatus* AF293 accession**	**Common name**	**Gene function**	***A. novofumigatus* orthologue**	**% ID**
AFUA_1G01550	zrfA	High affinity zinc ion transporter, putative [Source:UniProtKB/TrEMBL;Acc:Q4WKR5]	jgi-Aspnov1-378097-e_gw1.9.842.1	91.92
AFUA_1G05800	mkk2	MAP kinase kinase (Mkk2), putative [Source:UniProtKB/TrEMBL;Acc:Q4WJJ0]	jgi-Aspnov1-451544-estExt_Genemark1.C_4_t30113	90.68
AFUA_1G09280	ptcB	Protein phosphatase 2C, putative [Source:UniProtKB/TrEMBL;Acc:Q4WTH5]	jgi-Aspnov1-460512-fgenesh1_pm.4_#_814	91.44
AFUA_1G10880	pmcA	P-type calcium ATPase, putative [Source:UniProtKB/TrEMBL;Acc:Q4WT17]	jgi-Aspnov1-420834-gm1.5181_g	92.44
AFUA_1G14660	laeA	Regulator of secondary metabolism LaeA [Source:UniProtKB/TrEMBL;Acc:Q4WRY5]	jgi-Aspnov1-370504-e_gw1.4.2095.1	98.66
AFUA_1G15440	ags3	Alpha-1,3-glucan synthase Ags3 [Source:UniProtKB/TrEMBL;Acc:Q4WRQ8]	jgi-Aspnov1-447502-estExt_Genemark1.C_1_t20487	54.83
AFUA_1G16950	pig-a	Phosphatidylinositol:UDP-GlcNAc transferase subunit PIG-A [Source:UniProtKB/TrEMBL;Acc:Q4WRA7]	jgi-Aspnov1-405712-estExt_Genewise1.C_4_t10229	97.96
AFUA_2G01260	srbA	HLH transcription factor, putative [Source:UniProtKB/TrEMBL;Acc:Q4WIN1]	jgi-Aspnov1-455802-fgenesh1_pm.1_#_129	95.14
AFUA_2G07680	sidA	L-ornithine N5-oxygenase SidA [Source:UniProtKB/TrEMBL;Acc:E9QYP0]	jgi-Aspnov1-447204-estExt_Genemark1.C_1_t20174	96.01
AFUA_2G07770	rasB	Ras small monomeric GTPase RasB [Source:UniProtKB/TrEMBL;Acc:Q4X241]	jgi-Aspnov1-363087-e_gw1.1.2643.1	99.16
AFUA_2G08360	pyrG	Orotidine 5'-phosphate decarboxylase [Source:UniProtKB/Swiss-Prot;Acc:O13410]	jgi-Aspnov1-399623-estExt_Genewise1.C_1_t30405	97.81
AFUA_2G11270	ags2	Alpha-1,3-glucan synthase Ags2 [Source:UniProtKB/TrEMBL;Acc:Q4X143]	jgi-Aspnov1-447502-estExt_Genemark1.C_1_t20487	96.78
AFUA_2G12200	pkaC1	cAMP-dependent protein kinase catalytic subunit PkaC1 [Source:UniProtKB/TrEMBL;Acc:Q4X0V1]	jgi-Aspnov1-363483-e_gw1.1.1181.1	90.24
AFUA_2G12640	gprD	Integral membrane protein [Source:UniProtKB/TrEMBL;Acc:Q4X0Q7]	jgi-Aspnov1-456628-fgenesh1_pm.1_#_955	90.43
AFUA_2G17530	Melanin cluster, arb2	Conidial pigment biosynthesis oxidase Arb2 [Source:UniProtKB/TrEMBL;Acc:E9RBR0]	jgi-Aspnov1-501077-estExt_fgenesh1_pg.C_1_t40073	92.15
AFUA_2G17540	Melanin cluster, abr1	Conidial pigment biosynthesis oxidase Abr1/brown 1 [Source:UniProtKB/TrEMBL;Acc:Q4WZB4]	jgi-Aspnov1-417292-gm1.1639_g	88.61
AFUA_2G17550	Melanin cluster, ayg1	Conidial pigment biosynthesis protein Ayg1 [Source:UniProtKB/TrEMBL;Acc:Q4WZB3]	jgi-Aspnov1-438093-fgenesh1_kg.1_#_1767_#_Locus2445v1rpkm48.63	94.2
AFUA_2G17560	Melanin cluster, arp2	Conidial pigment biosynthesis 1,3,6,8-tetrahydroxynaphthalene reductase Arp2 [Source:UniProtKB/TrEMBL;Acc:E9QUT3]	jgi-Aspnov1-460884-fgenesh1_pm.5_#_30	49.22
AFUA_2G17580	Melanin cluster, arp1	Probable scytalone dehydratase [Source:UniProtKB/Swiss-Prot;Acc:O14434]	jgi-Aspnov1-457049-fgenesh1_pm.1_#_1376	95.15
AFUA_2G17600	Melanin cluster, alb1	Conidial pigment polyketide synthase PksP/Alb1 [Source:UniProtKB/TrEMBL;Acc:Q4WZA8]	jgi-Aspnov1-448094-estExt_Genemark1.C_1_t40143	94.78
AFUA_3G05650	orlA	Alpha,alpha-trehalose-phosphate synthase subunit Tps2, putative [Source:UniProtKB/TrEMBL;Acc:Q4WWF5]	jgi-Aspnov1-514040-estExt_fgenesh1_pm.C_6_t10212	97.88
AFUA_3G09820	dvrA	C2H2 transcription factor, putative [Source:UniProtKB/TrEMBL;Acc:Q4WXK4]	jgi-Aspnov1-462495-fgenesh1_pm.6_#_581	93.38
AFUA_3G11250	ace2	C2H2 transcription factor (Swi5), putative [Source:UniProtKB/TrEMBL;Acc:Q4WXZ7]	jgi-Aspnov1-423283-gm1.7630_g	87.45
AFUA_3G11970	pacC	pH-response transcription factor pacC/RIM101 [Source:UniProtKB/Swiss-Prot;Acc:Q4WY67]	jgi-Aspnov1-514555-estExt_fgenesh1_pm.C_6_t20262	89.44

AFUA_3G12690	glfA	Putative uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:Q4WYD9]	jgi-Aspnov1-446369-fgenesh1_kg.13_#_42_#_Locus6047v1rpkm11.26	74.92
AFUA_4G06820	ecm33	Protein ecm33 [Source:UniProtKB/Swiss-Prot;Acc:Q4WNS8]	jgi-Aspnov1-442983-fgenesh1_kg.5_#_827_#_Locus141v1rpkm963.23	87
AFUA_4G11800	Alp1	Alkaline protease 1 [Source:UniProtKB/Swiss-Prot;Acc:P28296]	jgi-Aspnov1-431992-fgenesh1_pg.5_#_295	94.29
AFUA_4G12470	cpcA	BZIP transcription factor CpcA [Source:UniProtKB/TrEMBL;Acc:E9QUZ5]	jgi-Aspnov1-461084-fgenesh1_pm.5_#_230	90.87
Afua_4g14770	Helvolic acid cluster	Protostadienol synthase A [Source:UniProtKB/Swiss-Prot;Acc:Q4WR16]	jgi-Aspnov1-501426-estExt_fgenesh1_pg.C_2_t10392	42.86
Afua_4g14780	Helvolic acid cluster	Cytochrome P450 monooxygenase, putative [Source:UniProtKB/TrEMBL;Acc:Q4WR17]	jgi-Aspnov1-458437-fgenesh1_pm.2_#_1359	47.7
Afua_4g14790	Helvolic acid cluster	Cytochrome P450 monooxygenase, putative [Source:UniProtKB/TrEMBL;Acc:Q4WR18]	jgi-Aspnov1-458437-fgenesh1_pm.2_#_1359	92.5
Afua_4g14800	Helvolic acid cluster	Short chain dehydrogenase, putative [Source:UniProtKB/TrEMBL;Acc:Q4WR19]	jgi-Aspnov1-439546-fgenesh1_kg.2_#_1427_#_Locus1689v1rpkm76.29	94
Afua_4g14810	Helvolic acid cluster	Cytochrome P450 monooxygenase, putative [Source:UniProtKB/TrEMBL;Acc:Q4WR20]	jgi-Aspnov1-458435-fgenesh1_pm.2_#_1357	85.6
Afua_4g14820	Helvolic acid cluster	Transferase family protein [Source:UniProtKB/TrEMBL;Acc:Q4WR21]	jgi-Aspnov1-439545-fgenesh1_kg.2_#_1426_#_Locus3005v1rpkm38.16	91
Afua_4g14830	Helvolic acid cluster	Cytochrome P450 monooxygenase, putative [Source:UniProtKB/TrEMBL;Acc:Q4WR22]	jgi-Aspnov1-458437-fgenesh1_pm.2_#_1359	47.59
Afua_4g14840	Helvolic acid cluster	Transferase family protein [Source:UniProtKB/TrEMBL;Acc:Q4WR23]	jgi-Aspnov1-439543-fgenesh1_kg.2_#_1424_#_Locus3740v1rpkm28.06	90
Afua_4g14850	Helvolic acid cluster	Extracellular 3-ketosteroid 1-dehydrogenase, putative [Source:UniProtKB/TrEMBL;Acc:Q4WR24]	jgi-Aspnov1-118200-CE118199_1683	91.79
AFUA_5G04170	hsp90	Heat shock protein 90 [Source:UniProtKB/Swiss-Prot;Acc:P40292]	jgi-Aspnov1-509883-estExt_fgenesh1_pm.C_2_t10392	99.27
AFUA_5G08570	pkaC2	cAMP-dependent protein kinase catalytic subunit, putative [Source:UniProtKB/TrEMBL;Acc:E9QXD5]	jgi-Aspnov1-100079-CE100078_981	95.71
AFUA_5G09240	cu/zn sod	Superoxide dismutase [Source:UniProtKB/Swiss-Prot;]	jgi-Aspnov1-501817-estExt_fgenesh1_pg.C_2_t20321	97.33
AFUA_5G09360	calA	Serine/threonine-protein phosphatase 2B catalytic subunit [Source:UniProtKB/Swiss-Prot;Acc:Q4WUR1]	jgi-Aspnov1-457892-fgenesh1_pm.2_#_814	97.59
AFUA_5G09580	rodA	Hydrophobin [Source:UniProtKB/Swiss-Prot;Acc:P41746]	jgi-Aspnov1-402332-estExt_Genewise1.C_2_t40215	96.43
AFUA_5G10760	mnt1	Alpha-1,2-mannosyltransferase (Kre2), putative [Source:UniProtKB/TrEMBL;Acc:Q4WV44]	jgi-Aspnov1-108431-CE108430_6570	94
AFUA_5G11230	rasA	RAS small monomeric GTPase RasA [Source:UniProtKB/TrEMBL;Acc:E9QX28]	jgi-Aspnov1-109419-CE109418_6451	99.53
AFUA_5G13300	pep1	Aspartic protease pep1 [Source:UniProtKB/Swiss-Prot;Acc:P41748]	jgi-Aspnov1-363785-e_gw1.2.2138.1	95.19
AFUA_6G04820	pabaA	Para-aminobenzoate synthase PabaA [Source:UniProtKB/TrEMBL;Acc:Q4WDI0]	jgi-Aspnov1-450180-estExt_Genemark1.C_3_t20429	88.47
AFUA_6G09570	Gliotoxin cluster	Putative uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:Q4WMK6]	jgi-Aspnov1-508952-estExt_fgenesh1_pm.C_1_t20245	30.19
AFUA_6G09580	Gliotoxin cluster	C6 finger domain protein, putative [Source:UniProtKB/TrEMBL;Acc:Q4WMK5]	jgi-Aspnov1-511168-estExt_fgenesh1_pm.C_3_t20002	82.02
AFUA_6G09590	Gliotoxin cluster	Zinc alcohol dehydrogenase, putative [Source:UniProtKB/TrEMBL;Acc:Q4WMK4]	jgi-Aspnov1-511167-estExt_fgenesh1_pm.C_3_t20001	84.89
AFUA_6G09600	Gliotoxin cluster	Zinc metallopeptidase, putative [Source:UniProtKB/TrEMBL;Acc:Q4WMK3]	jgi-Aspnov1-502750-estExt_fgenesh1_pg.C_3_t20015	90.57

AFUA_6G09610	Gliotoxin cluster	Nonribosomal peptide syntethase 9 [Source:UniProtKB/Swiss-Prot;Acc:Q4WMK2]	jgi-Aspnov1-404104-estExt_Genewise1.C_3_t30145	77.01
AFUA_6G09620	Gliotoxin cluster	Putative uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:Q4WMK1]	jgi-Aspnov1-386801-estExt_Genewise1Plus.C_3_t30133	84.72
AFUA_6G09630	Gliotoxin cluster	C6 finger domain protein GliZ [Source:UniProtKB/TrEMBL;Acc:Q4WMK0]	jgi-Aspnov1-366288-e_gw1.3.282.1	83.75
AFUA_6G09640	Gliotoxin cluster	Aminotransferase GliI [Source:UniProtKB/TrEMBL;Acc:Q4WMJ9]	jgi-Aspnov1-458936-fgenesh1_pm.3_#_494	87.07
AFUA_6G09650	Gliotoxin cluster	Membrane dipeptidase GliJ [Source:UniProtKB/TrEMBL;Acc:Q4WMJ8]	jgi-Aspnov1-367031-e_gw1.3.965.1	91.24
AFUA_6G09660	Gliotoxin cluster	Nonribosomal peptide synthetase 10 [Source:UniProtKB/Swiss-Prot;Acc:Q4WMJ7]	jgi-Aspnov1-511160-estExt_fgenesh1_pm.C_3_t10488	92.06
AFUA_6G09670	Gliotoxin cluster	Cytochrome P450 oxidoreductase GliC [Source:UniProtKB/TrEMBL;Acc:E9RCR4]	jgi-Aspnov1-481241-MIX15903_10_44	92.59
AFUA_6G09680	Gliotoxin cluster	O-methyltransferase GliM [Source:UniProtKB/TrEMBL;Acc:Q4WMJ5]	jgi-Aspnov1-368520-e_gw1.3.829.1	93.5
AFUA_6G09690	Gliotoxin cluster	Glutathione S-transferase GliG [Source:UniProtKB/TrEMBL;Acc:A4GYZ0]	jgi-Aspnov1-458931-fgenesh1_pm.3_#_489	96.25
AFUA_6G09700	Gliotoxin cluster	Gliotoxin biosynthesis protein GliK [Source:UniProtKB/TrEMBL;Acc:E9R9Y3]	jgi-Aspnov1-429762-fgenesh1_pg.3_#_503	90.33
AFUA_6G09710	Gliotoxin cluster	MFS gliotoxin efflux transporter GliA [Source:UniProtKB/TrEMBL;Acc:E9R876]	jgi-Aspnov1-367489-e_gw1.3.2721.1	94.1
AFUA_6G09720	Gliotoxin cluster	Methyltransferase GliN [Source:UniProtKB/TrEMBL;Acc:Q4WMJ1]	jgi-Aspnov1-502743-estExt_fgenesh1_pg.C_3_t20001	84.04
AFUA_6G09730	Gliotoxin cluster	Cytochrome P450 oxidoreductase GliF [Source:UniProtKB/TrEMBL;Acc:Q4WMJ0]	jgi-Aspnov1-458927-fgenesh1_pm.3_#_485	95.44
AFUA_6G09740	Gliotoxin cluster	Thioredoxin reductase GliT [Source:UniProtKB/TrEMBL;Acc:E9RAH5]	jgi-Aspnov1-511158-estExt_fgenesh1_pm.C_3_t10480	91.92
AFUA_6G09745	Gliotoxin cluster	Putative uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:Q4WMI8]	jgi-Aspnov1-386790-estExt_Genewise1Plus.C_3_t30100	89.84
AFUA_6G10240	fos-1 (tcsA)	Sensor histidine kinase/response regulator Fos-1/TcsA [Source:UniProtKB/TrEMBL;Acc:Q4WMD9]	jgi-Aspnov1-429720-fgenesh1_pg.3_#_461	93.36
AFUA_6G11390	gel2	1,3-beta-glucanosyltransferase gel2 [Source:UniProtKB/Swiss-Prot;Acc:P0C954]	jgi-Aspnov1-398517-estExt_Genewise1.C_1_t10227	48.38
AFUA_7G04800	gprC	Putative uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:Q4WGE9]	jgi-Aspnov1-462935-fgenesh1_pm.7_#_220	94.32
AFUA_8G00170	Fumitremorgin cluster	Nonribosomal peptide synthetase 13 [Source:UniProtKB/Swiss-Prot;Acc:Q4WAW3]	jgi-Aspnov1-507323-estExt_fgenesh1_pg.C_90140	34.91
AFUA_8G00190	Fumitremorgin cluster	Cytochrome P450, putative [Source:UniProtKB/TrEMBL;Acc:Q4WAW5]	jgi-Aspnov1-510751-estExt_fgenesh1_pm.C_3_t10008	64.49
AFUA_8G00200	Fumitremorgin cluster	O-methyltransferase, putative [Source:UniProtKB/TrEMBL;Acc:Q4WAW6]	jgi-Aspnov1-449391-estExt_Genemark1.C_3_t10007	71.39
AFUA_8G00210	Fumitremorgin cluster	Dimethylallyl tryptophan synthase FtmPT1 [Source:UniProtKB/TrEMBL;Acc:Q4WAW7]	jgi-Aspnov1-396938-estExt_Genewise1Plus.C_10_t10291	35.94
AFUA_8G00220	Fumitremorgin cluster	Cytochrome P450, putative [Source:UniProtKB/TrEMBL;Acc:Q4WAW8]	jgi-Aspnov1-479619-MIX14281_2_12	44.02
AFUA_8G00230	Fumitremorgin cluster	Phytanoyl-CoA dioxygenase family protein [Source:UniProtKB/TrEMBL;Acc:Q4WAW9]	jgi-Aspnov1-404130-estExt_Genewise1.C_3_t30177	33.3
AFUA_8G00240	Fumitremorgin cluster	Cytochrome P450 monooxygenase, putative [Source:UniProtKB/TrEMBL;Acc:Q4WAX0]	jgi-Aspnov1-419348-gm1.3695_g	42.5
AFUA_8G00250	Fumitremorgin cluster	Dimethylallyl tryptophan synthase, putative [Source:UniProtKB/TrEMBL;Acc:Q4WAX1]	jgi-Aspnov1-463437-fgenesh1_pm.8_#_19	57.76
AFUA_8G00260	Fumitremorgin cluster	F-box domain and ankyrin repeat protein [Source:UniProtKB/TrEMBL;Acc:Q4WAX2]	jgi-Aspnov1-516072-estExt_fgenesh1_pm.C_100234	27.92
AFUA_8G00370	Fumagillin cluster	Polyketide synthase, putative [Source:UniProtKB/TrEMBL;Acc:Q4WAY3]	jgi-Aspnov1-424178-gm1.8525_g	84.34

AFUA_8G00380	Fumagillin cluster	DltD N-terminal domain protein [Source:UniProtKB/TrEMBL;Acc:Q4WAY4]	jgi-Aspnov1-412667-estExt_Genewise1.C_8_t10092	96.92
AFUA_8G00390	Fumagillin cluster	O-methyltransferase, putative [Source:UniProtKB/TrEMBL;Acc:Q4WAY5]	jgi-Aspnov1-463456-fgenesh1_pm.8_#_38	96.77
AFUA_8G00400	Fumagillin cluster	Putative uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:Q4WAY6]	jgi-Aspnov1-463456-fgenesh1_pm.8_#_38	85.06
AFUA_8G00410	Fumagillin cluster	Methionine aminopeptidase 2-1 [Source:UniProtKB/Swiss-Prot;Acc:Q4WAY7]	jgi-Aspnov1-463457-fgenesh1_pm.8_#_39	91.48
AFUA_8G00420	Fumagillin cluster	C6 finger transcription factor, putative [Source:UniProtKB/TrEMBL;Acc:Q4WAY8]	jgi-Aspnov1-434405-fgenesh1_pg.8_#_51	86.95
AFUA_8G00430	Fumagillin cluster	Putative uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:Q4WAY9]	jgi-Aspnov1-412676-estExt_Genewise1.C_8_t10101	95.54
AFUA_8G00440	Fumagillin cluster	Steroid monooxygenase, putative [Source:UniProtKB/TrEMBL;Acc:Q4WAZ0]	jgi-Aspnov1-434407-fgenesh1_pg.8_#_53	87.4
AFUA_8G00460	Fumagillin cluster	Methionine aminopeptidase [Source:UniProtKB/TrEMBL;Acc:Q4WAZ1]	jgi-Aspnov1-463462-fgenesh1_pm.8_#_44	96.22
AFUA_8G00470	Fumagillin cluster	Putative uncharacterized protein [Source:UniProtKB/TrEMBL;Acc:Q4WAZ2]	jgi-Aspnov1-395381-estExt_Genewise1Plus.C_8_t10105	85.82
AFUA_8G00480	Fumagillin cluster	Phytanoyl-CoA dioxygenase family protein [Source:UniProtKB/TrEMBL;Acc:Q4WAZ3]	jgi-Aspnov1-445026-fgenesh1_kg.8_#_46_#_Locus3448v1rpkm31.50	94.01
AFUA_8G00490	Fumagillin cluster	PKS-like enzyme, putative [Source:UniProtKB/TrEMBL;Acc:Q4WAZ4]	jgi-Aspnov1-445027-fgenesh1_kg.8_#_47_#_Locus11024v1rpkm1.79	72.1
AFUA_8G00500	Fumagillin cluster	Acetate-CoA ligase, putative [Source:UniProtKB/TrEMBL;Acc:Q4WAZ5]	jgi-Aspnov1-412687-estExt_Genewise1.C_8_t10112	92.09
AFUA_8G00510	Fumagillin cluster	Cytochrome P450 oxidoreductase OrdA-like, putative [Source:UniProtKB/TrEMBL;Acc:Q4WAZ6]	jgi-Aspnov1-412689-estExt_Genewise1.C_8_t10114	94.03
AFUA_8G00520	Fumagillin cluster	Integral membrane protein [Source:UniProtKB/TrEMBL;Acc:Q4WAZ7]	jgi-Aspnov1-377876-e_gw1.8.1481.1	91.23
AFUA_8G02750	cgrA	rRNA-processing protein cgrA [Source:UniProtKB/Swiss-Prot;Acc:Q9HEQ8]	jgi-Aspnov1-454324-estExt_Genemark1.C_8_t10302	95

Table S4 Table of the terrequinone proteins from *A. nidulans* and the BLASTP hits found in *A. steynii*.

A. Nidulans terrequinone proteins	Hit protein ID in *A. steynii*	% Identity	% Coverage
TdiA - ABU51602.1	365047	71.7	101
tdiB - ABU51603.1	415228	56.2	92
tdiC - ABU51604.1	429405	52.7	88.6
tdiD - ABU51605.1	479193	69.5	96.1
tdiE - ABU51606.1	365428	56.1	97.7

Table S5

Part 1 Overview of chlorinating enzymes identified from literature. The sequence of each of these proteins were used to search for similar proteins in *A. campestris, A. candidus* and *A. taichungensis* using BLASTP comparison, but no hits were found.

Protein	Description	Reference	GenBank	*A. campestris*	*A. candidus*	*A. taichungensis*
CmaB	a chlorinating non-haem iron enzyme from Pseudomonas syringae	[1]	AAC46036.1	No hits	No hits	No hits
PrnA	a flavin dependent tryptophan halogenase from *Pseudomonas fluorescens*	[2]	AAB97504.1	No hits	No hits	No hits
PtaM	a flavin dependent halogenase from *Pestalotiopsis fici*	[3]	AGO59046.1	No hits	No hits	No hits
Thr3	an iron non heme alpha ketoglutarate dependent halogenase from Streptomyces sp. OH-5093	[4]	CCF23457.1	No hits	No hits	No hits

Part 2 Relevant InterPro domains identified by comparing the proteins listed in Suppl. Table 6 to the InterPro database using InterProScan 5 [5] and an additional word search of the database [6]. The identified InterPro domains were searched for in the annotated *A. campestris, A. candidus* and *A. taichungensis* genomes including the number of hits.

IPR ID	Description	Hits *A. campestris*	Hits *A. candidus*	Hits *A. taichungensis*
IPR000028	Chloroperoxidase Heme dependent	4	4	4
IPR001568	Ribonuclease T2	3	2	2
IPR008775	Phytanoyl-CoA dioxygenase	5	5	4
IPR010092	chlorinating enzyme FE(II)nonheme halogenase	0	0	0
IPR006905	tryptophan halogenase	0	0	0
IPR002747	SAM depdent chlorinase/fluorinase	0	0	0
IPR016119	bromoperoxidase/chloroperoxidase C-terminal	0	0	0

Part 3 Overview of the potential chlorinating proteins in the chlorflavonin candidate cluster in *A. campestris* including the best BLASTP hit in NCBI nr database and the identified InterPro IDs found using InterPro Scan 5 [5] on the protein sequences.

Protein ID	BLAST hit	InterPro Scan hit
286063	Hypothetical protein [Solirubrobacterales bacterium URHD0059] 93% coverage and 36% identity	IPR029039 – flavoprotein-like IPR005025 – NADPH-dependent FMN reductase-like
277538	Related to scytalone dehydratase [Fusarium fujikuroi IMI 58289], 100% coverage and 53% identity	IPR004235 – Scytalone dehydratase

331187	Peptidase S15/CocE/NonD, C-terminal [Penicillium expansum] 97% coverage and 41% Identity	IPR011008 Dimeric alpha-beta barrel
3988	Hypothetical protein HIM_08269 [Hirsutella minnesotensis 3608] 98% coverage and 53% Identity 3-hydroxybenzoate 6-hydroxylase 1 [Tolypocladium ophioglossoides CBS 100239] 98% coverage and 53% Identity	IPR006076 –FAD dependent oxidoreductase IPR003042 – Aromatic-ring hydroxylase IPR023753 – FAD/NAD(P) binding domain IPR002938 – FAD-binding domain

1. Vaillancourt FH, Yeh E, Vosburg D a, O'Connor SE, Walsh CT. Cryptic chlorination by a non-haem iron enzyme during cyclopropyl amino acid biosyn. Nature. 2005;436:1191–4. http://www.nature.com/nature/journal/v436/n7054/pdf/nature03797.pdf. Accessed 3 Nov 2015.
2. Kirner S, Hammer PE, Hill DS, Altmann A, Fischer I, Weislo LJ, et al. Functions Encoded by Pyrrolnitrin Biosynthetic Genes from Pseudomonas fluorescens. J Bacteriol. 1998;180:1939–43. http://jb.asm.org.proxy.findit.dtu.dk/content/180/7/1939. Accessed 3 Nov 2015.
3. Xu X, Liu L, Zhang F, Wang W, Li J, Guo L, et al. Identification of the first diphenyl ether gene cluster for pestheic acid biosynthesis in plant endophyte Pestalotiopsis fici. Chembiochem. 2014;15:284–92. doi:10.1002/cbic.201300626.
4. Fullone MR, Paiardini A, Miele R, Marsango S, Gross DC, Omura S, et al. Insight into the structure-function relationship of the nonheme iron halogenases involved in the biosynthesis of 4-chlorothreonine --Thr3 from Streptomyces sp. OH-5093 and SyrB2 from Pseudomonas syringae pv. syringae B301DR. FEBS J. 2012;279:4269–82. doi:10.1111/febs.12017.
5. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30:1236–40. doi:10.1093/bioinformatics/btu031.
6. Mitchell A, Chang H-Y, Daugherty L, Fraser M, Hunter S, Lopez R, et al. The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Res. 2014;43:D213-221. doi:10.1093/nar/gku1243.

B Supplementary material section 3.2 – Manuscript II

Following is the supplementary material for the article presented in Chapter 3 section 3.2 'Friends and foes – A comparative genomics study of 23 *Aspergillus* species from section Flavi'.

Table B.1 Growth section *Flavi* quantitatively. Growth analysis of 23 *Flavi* species plus 8 additional species on 35 different growth media, quantitated by growth from 0-10, normalized based on growth on 1% glucose

Supplementary Table B.1 can be found using the following link:
https://files.dtu.dk/u/qpZ1Lj8WdSS4lnUs/TableB1_GrowthQuatitatively.xlsx?l

Table B.2 CAZy content in section *Flavi*. Overview of the CAZy content and plant degradation related CAZy content per species and per CAZy families.

Supplementary Table B.2 can be found using the following link:
https://files.dtu.dk/u/4b84FHQ_OB_zvzgW/TableB2_CAZy_analysis.xlsx?l

Table B.3 Secondary metabolite gene clusters section *Flavi*. Long format table with all the predicted clusters in the species and the cluster family they belong to. Column 1 - Species, column 2 - JGI protein id of the predicted backbone, column 3 - cluster family number.

Supplementary Table B.3 can be found using the following link:
https://files.dtu.dk/u/mS3dxwyt3c-LLd3F/TableB3_SecMetFamilies.csv?l

Table B.4 Chemical analysis of known compounds produced by *Flavi* species after growth on CYA for 7 days at 30°C.

Supplementary Table B.4 can be found using the following link:
https://files.dtu.dk/u/5YhxWHAIcVH3iHFa/TableB4_ChemicalDataFlavi.xlsx?l

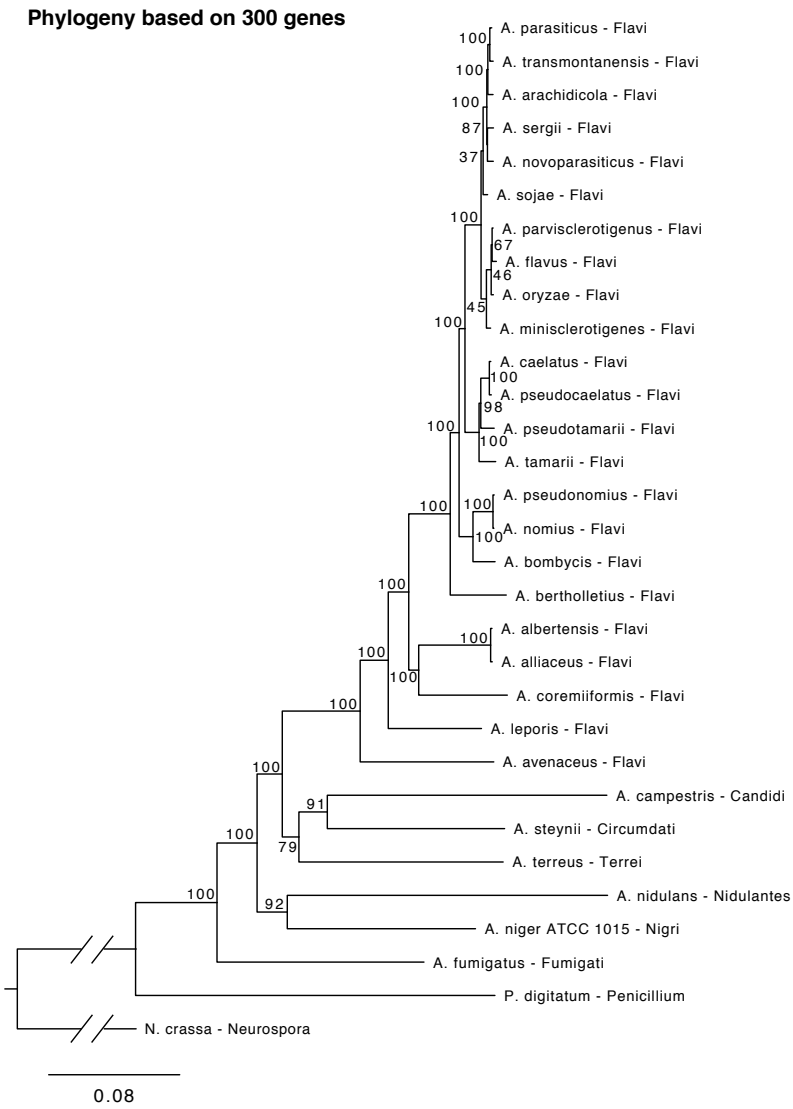


Figure B.1 Phylogenetic tree based on 300 monocore genes. Phylogenetic tree constructed using RAxML [119], MUSCLE [120], and Gblock [121] based on 300 monocore genes.

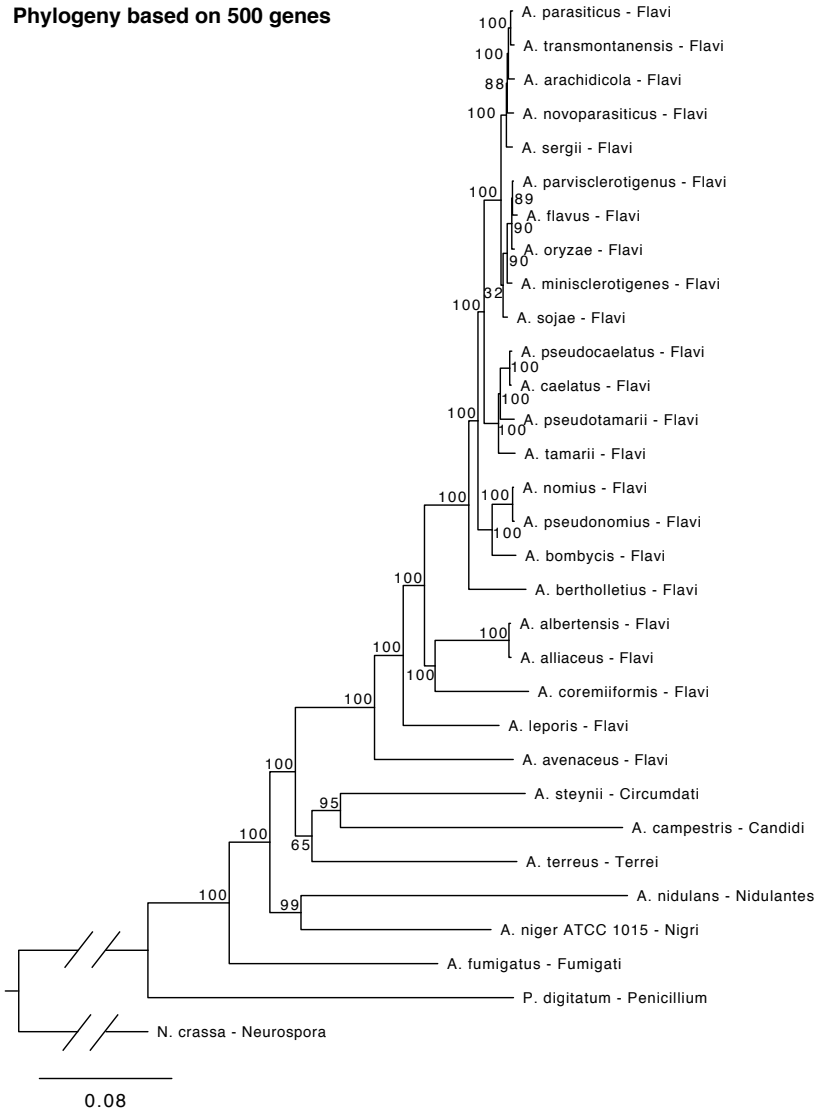


Figure B.2 Phylogenetic tree based on 500 monocore genes. Phylogenetic tree constructed using RAxML [119], MUSCLE [120], and Gblock [121] based on 200 monocore genes.

Supplementary Figures B.3 can be found using the following link:
https://files.dtu.dk/u/xixzBFOscwjULHMG/FigureB3?l

Figure B.3 Phylogenetic trees constructed using RAxML [119] and MUSCLE [120] of 18 single gene tress from monocore families.

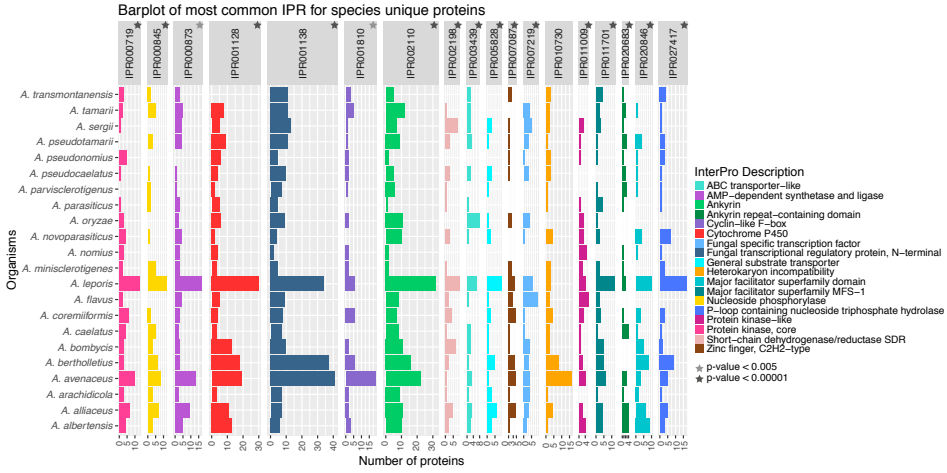


Figure B.4 Most common InterPro domains in species unique proteins. Bar plot showing the number of species unique proteins with an InterPro domain per species, shown for the most common InterPro annotations [102]. Light grey star indicates p-values below 0.005 and dark grey star indicates p-value below 0.00001 of enrichment in the species unique genes for the specific functional domain.

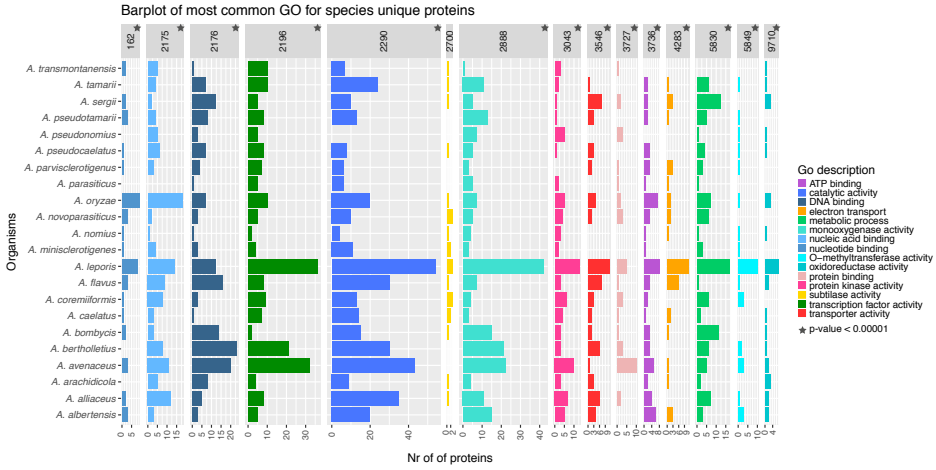


Figure B.5 Most common GO domains in species unique proteins. Bar plot showing the number of species unique proteins with a GO domain per species, shown for the most common GO annotations [99]. Dark grey star indicates p-value below 0.00001 of enrichment in the species unique genes for the specific functional domain.

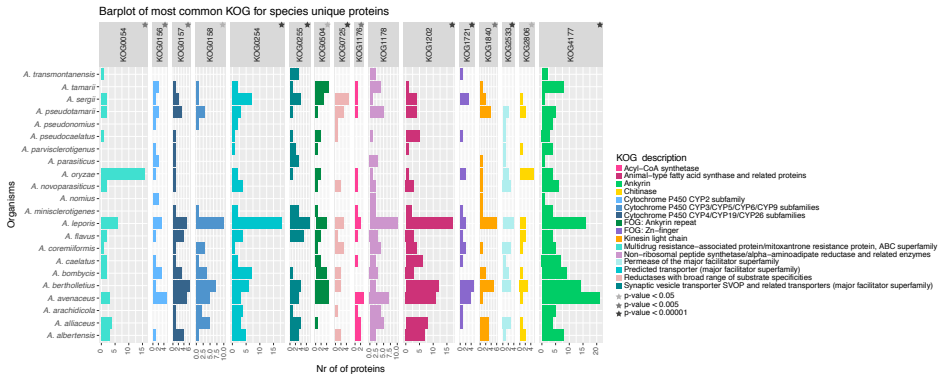


Figure B.6 Most common KOG domains in species unique proteins. Bar plot showing the number of species unique proteins with an KOG domain per species, shown for the most common KOG annotations [100]. P-values indicates enrichment of a certain KOG term in the species unique genes for the specific functional domain. The p-values are indicated by stars of light to dark grey the lightest star indicates p-values below 0.05, followed by p-values below 0.005 and the darkest stars indicates p-value below 0.00001.

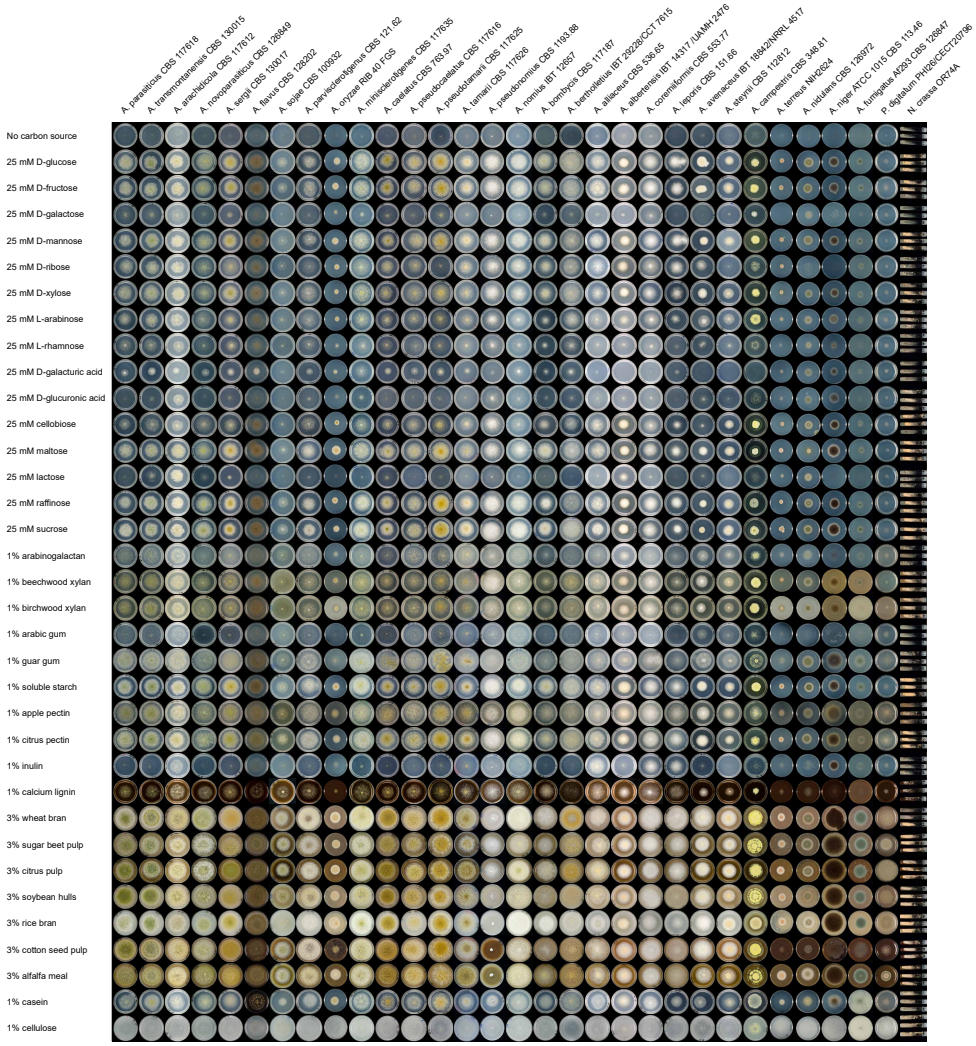


Figure B.7 Growth section *Flavi*. Growth analysis of 23 *Flavi* species plus 8 additional species on 35 different growth media.

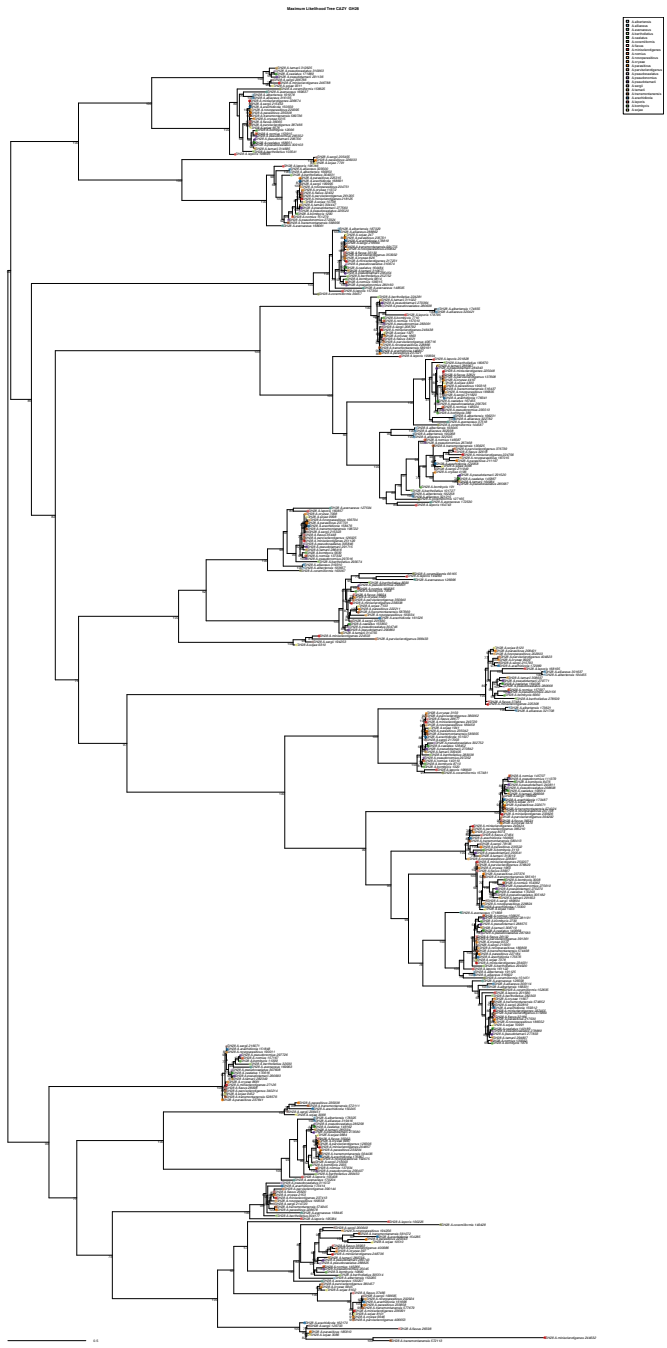


Figure B.8 Phylogenetic tree of GH28. Phylogenetic tree of all proteins assigned to the GH28 CAZy catagory. The GH28 family consits of polygalacturonase. Alignment of the members of GH28 CAZy family found in all section *Flavi* species was created using clustalo. The ML phylogenetic tree was created using the ape package in R.

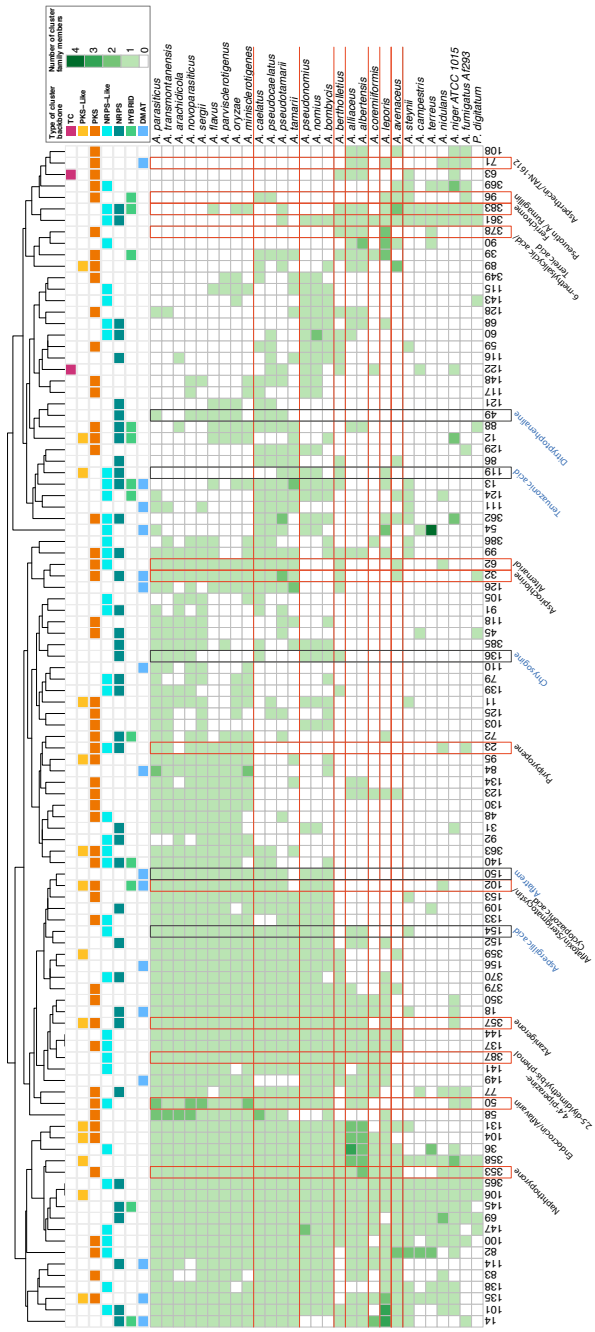


Figure B.9 Heatmap cluster families. Cluster families with members in at least 5 species are illustrated by a heatmap. The top rows indicate the backbone enzymes found within the cluster family. Compounds with similar clusters are added from the dereplication using MIBiG (marked by orange boxes and black text) in addition to manually curated compounds (marked by black boxes and blue text). Aspergillic acid [122], aflatrem [123], chrysogine [124], tenuazonic acid [125], dityptophenaline [126].

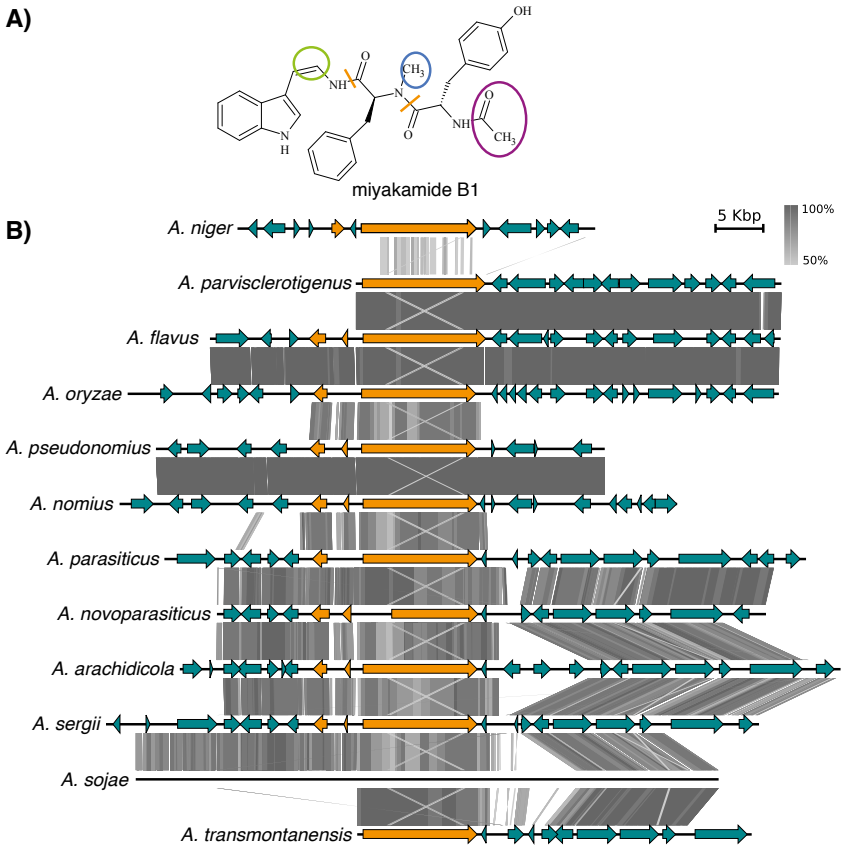


Figure B.10 Miyakamide and putative clusters. A) Miyakamide B1 showing the three amino acid parts (organge line), the acetylation (purple circle), the decarboxylation (green circle) and the N-methylation (blue circle). B) Syntenic plot of the putative miyakamide cluster family plus surrounding genes. The syntney plot was generated using EasyFig [127] with the minimum length 50 bp and the minimum identity to 50%. The genes potentially involved in miyakamide production are marked by orange.

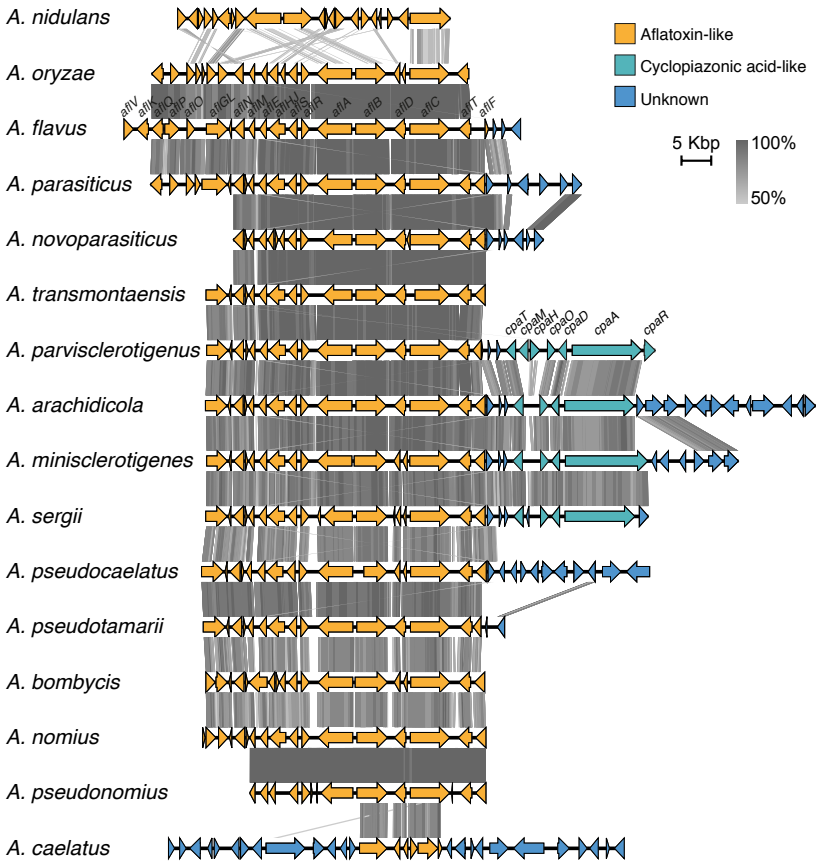


Figure B.11 Synteny of aflatoxin cluster family. Syntenic plot of the predicted clusters belonging to the 'sterigmatocystin - aflatoxin - cyclopiazonic-acid' cluster family. The syntney plot was generated using EasyFig [127] with the minimum length 50 bp and the minimum identity to 50%. The genes are color coded based on known pathways; aflatoxin (orange), cyclopiazonic-acid (turquoise) and unknown (blue).

```
CLUSTAL O(1.2.4) multiple sequence alignment


A.flavus_36747                  mlcfrsvdglstslaevsrlpssllvlsrilqickskdattvrtrlnrhtrhgplssfhs   60
A.bombycis_5312                 ------------------------------------------------------------   0
A.pseudocaelatus_279719         ------------------------------------------------------------   0
A.pseudotamarii_276010          ------------------------------------------------------------   0
A.nomius_129446                 ------------------------------------------------------------   0
A.pseudonomius_274161           ------------------------------------------------------------   0
A.sergii_195135                 ------------------------------------------------------------   0
A.minisclerotigenes_217119      ------------------------------------------------------------   0
A.transmontanensis_544984       ------------------------------------------------------------   0
A.oryzae_5677                   ------------------------------------------------------------   0
A.parvisclerotigenus_400163     ------------------------------------------------------------   0
A.arachidicola_161418           ------------------------------------------------------------   0
A.parasiticus_236245            ------------------------------------------------------------   0
A.novoparasiticus_191948        ------------------------------------------------------------   0


A.flavus_36747                  iepirpspikmtlpnkaalvglahtlseqvkrylvtadetkspedhklciegertpsste   120
A.bombycis_5312                 ----------malpnktalvglanalseqvksylattdgtelskdq-----svgspsste   45
A.pseudocaelatus_279719         ----------maqpnkaalvglantlselvkryvgtadetklsedhthcmesvrspsste   50
A.pseudotamarii_276010          ----------maqpnkaalvglantlseqvkrylgtadgaklsevdthcmdgvrspsste   50
A.nomius_129446                 ----------malpdkaalvglanalseqvkrylaiadetdlpnghkhciegannpssae   50
A.pseudonomius_274161           ----------malpdkaalvglanalseqvkrylaiadetdlpnghkhciegannpssae   50
A.sergii_195135                 ----------malpnkaalmglantlseqvrrylatadetkspedhklciesvrspsste   50
A.minisclerotigenes_217119      ----------malpnkaalvglahtlsdqvkrylvtaddtkspedhklcierertpsste   50
A.transmontanensis_544984       ----------malpnkaalvglantlseqvkcylatagetkspedhkvcieserapsste   50
A.oryzae_5677                   ----------malpnkaalvglahtlseqvkhylvtadetkspedhklciereitpsste   50
A.parvisclerotigenus_400163     ----------malpnkaalvglahtlseqvkhylvtadetkspedhklciereitpsste   50
A.arachidicola_161418           ----------malpnkaalvglantlseqvqryqatagetkspedhklciesertps-te   49
A.parasiticus_236245            ----------malpskaalvglantlseqvkrylatagetkspedhklciesertpssne   50
A.novoparasiticus_191948        ----------malpnkaallglantlseqvkrylatagetkspedhklciesernssste   50
                                          *: *.*:**:***::**: *: *   :. :.  : .          *  *

A.flavus_36747                  haqaweivrtcdrigslvhgpvpwllsnalshldsaclaaatqlnlqdiivdgpsptsld   180
A.bombycis_5312                 haqaweivrtcdrigslihgpvpwllsnalshldsaclaaavqlnlqdiivdgpsptsld   105
A.pseudocaelatus_279719         yaqaweivrtcdrisslihgpvpwllsnalshldsaclaaatqlnlqdiivdgpsptslk   110
A.pseudotamarii_276010          haqaweivrtcdrisslihgpvpwllsnalshldsaclaaatqlnlqdiivdgpsptslk   110
A.nomius_129446                 haqaweivrtcdrigslihgpvpwllsnalshldsaclaaatqlnlqdvivdgasptsld   110
A.pseudonomius_274161           haqaweivrtcdrigslihgpvpwllsnalshldsaclaaatqlnlqdvivdgasptsld   110
A.sergii_195135                 haqaweivrtcdrigslvhgpvpwllsnalshldsaclaaatqlnlqdiivdgpsptsld   110
A.minisclerotigenes_217119      haqaweivrtcdrigslvhgpvpwllsnalshldsaclaaatqlnlqdiivdgpspasld   110
A.transmontanensis_544984       haqaweivrtcdrigslvhgpvpwllsnalshldsaclaaathlnlqdiivdgpsptsld   110
A.oryzae_5677                   haqaweivrtcdrigslvhgpvpwllsnalshldsaclaaatqlnlqdiivdgpsptsld   110
A.parvisclerotigenus_400163     haqaweivrtcdrigslvhgpvpwllsnalshldsaclaaatqlnlqdiivdgpsptsld   110
A.arachidicola_161418           haqaweivrtcdrigslvhgpvpwllsnalshldsaclaaathlnlqdiivdgpsptsld   109
A.parasiticus_236245            haqaweivrtcdrigslvhgpvpwllsnalshldsaclaaathlnlqdiivdgpsptsld   110
A.novoparasiticus_191948        haqaweivrtcdrigslvhgpvpwllsnalshldsaclaaathlnlqdiivdgpsptsld   110
                                :************.**:**********************.:*****:**** **:**.

A.flavus_36747                  tivtatgvsedll-------------------------------------------------   193
A.bombycis_5312                 tivvatgvsedllrrilrgcaqrfifeevapgqyahtdaskmlrvtgihalvgfsrphqa   165
A.pseudocaelatus_279719         tivaatgvsedllrrilrgcaqrfifeevaadqfahtdaskmlcvtgihalvgfscpdqr   170
A.pseudotamarii_276010          tivaatgvsedllrrilrgcaqrlifeevaadqyahtdaskmlcvtgihalvgfscphqr   170
A.nomius_129446                 tivaatgvskdllrrilrgcaqrfifeevapdqyahtdasemlrvkgihalvgfs-----   165
A.pseudonomius_274161           tivaatgvskdllrrilrgcaqrfifeevapdqyahtdaskmlrvkgihalvgfs-----   165
A.sergii_195135                 tivaatgvsddllrrilrgcaqrfifeevapdqyahtdaskmlrvtgihalvgfscphqg   170
A.minisclerotigenes_217119      tivaatgvsedllrrilrgcaqrfifeevapdqyahtdaskmlrvtgihalvgfsc----   166
A.transmontanensis_544984       tivaatgvsedllrrilrgcaqrfifeevapdqyahtdaskmlrvtgihalvgfsc----   166
A.oryzae_5677                   tivaatgvsedllrrilrgcaqrfifeevapdqyahtdaskmlrvtgihalvgfscphqg   170
A.parvisclerotigenus_400163     tivaatgvsedllrrilrgcaqrfifeevapdqyahtdaskmlrvtgihalvgfscphqg   170
A.arachidicola_161418           tivaatgvsedllrrilrgcaqrfifeevapdqyahtdaskmlrvgihalvgfscphqg   169
A.parasiticus_236245            tivaatgvsedllrrilrgcaqrfifeevapdqyahtdaskmlrvtgihalvgfsc----   166
A.novoparasiticus_191948        tivaatgvsedllrrilrgcaqrfifeevapdqyahtdaskmlrvtgihalvgfscphqg   170
                                ***.*****.***

A.flavus_36747                  ----------------------------------------ptkglfdyystvdevrgrrf   213
A.bombycis_5312                 yrpscdevmrsgayfseflqqtkgnppswnvpspfslafdpakglfdyyntvdevrgrrf   225
A.pseudocaelatus_279719         drvscdevmrsgacfsdflqqtkgnppswnvpspfslafdpakglfdyysmvdevrgrrf   230
A.pseudotamarii_276010          drvscdevmhsgacfsdflqqtkgnppswnvpspfslafdpakglfdyysmvdesrgrrf   230
A.nomius_129446                 ----cdevmrsgacfsdflqqtkgnppswnvpspfslafdpakglfdyystvdevrgrrf   221
A.pseudonomius_274161           ----cdevmrsgacfsdflqqtkgnppswnvpspfslafdpakglfdyystvdevrgrrf   221
A.sergii_195135                 drvscdevmrsgacfsdflqqtkgnpprwnvpspfslafdptkglfdyystvdevrgrrf   230
A.minisclerotigenes_217119      -----devmrsgayfsdflqqtkdnppswnvpspfslafdptkglfdyystvdevrgrrf   221
A.transmontanensis_544984       -----devmrsgayfadflqqtkgkppswnvpspfslafdptkglfdyystvdevrgrrf   221
A.oryzae_5677                   drvrcdevmrsgayfsdflqqtkgkppswnvpspfslafdptkglfdyystvdevrgrrf   230
A.parvisclerotigenus_400163     drvrcdevmrsgayfsdflqqtkgkppswnvpspfslafdptkglfdyystvdevrgrrf   230
A.arachidicola_161418           drvrcdevmrsgayfsdflqqtkgkppswnvpspfslafdptkglfdyystvdevrgrrf   229
A.parasiticus_236245            -----devmrsgasfsdflqqtkgkppswnvpspfslafdptkglfdyystvdevrgrrf   221
A.novoparasiticus_191948        drvrcdevmrsgayfsdflqqtkgkppswnvpspfslafdptkglfdyystvdevrgrrf   230
                                                                 *:*******. *** *****
```

```
A.flavus_36747                  ----------------------------------------ptkglfdyystvdevrgrrf   213
A.bombycis_5312                 yrpscdevmrsgayfseflqqtkgnppswnvpspfslafdpakglfdyyntvdevrgrrf   225
A.pseudocaelatus_279719         drvscdevmrsgacfsdflqqtkgnppswnvpspfslafdpakglfdyysmvdevrgrrf   230
A.pseudotamarii_276010          drvscdevmhsgacfsdflqqtkgnppswnvpspfslafdpakglfdyysmvdesrgrrf   230
A.nomius_129446                 ----cdevmrsgacfsdflqqtkgnppswnvpspfslafdpakglfdyystvdevrgrrf   221
A.pseudonomius_274161           ----cdevmrsgacfsdflqqtkgnppswnvpspfslafdpakglfdyystvdevrgrrf   221
A.sergii_195135                 drvscdevmrsgacfsdflqqtkgnpprwnvpspfslafdptkglfdyystvdevrgrrf   230
A.minisclerotigenes_217119      -----devmrsgayfsdflqqtkdnppswnvpspfslafdptkglfdyystvdevrgrrf   221
A.transmontanensis_544984       -----devmrsgayfadflqqtkgkppswnvpspfslafdptkglfdyystvdevrgrrf   221
A.oryzae_5677                   drvrcdevmrsgayfsdflqqtkgkppswnvpspfslafdptkglfdyystvdevrgrrf   230
A.parvisclerotigenus_400163     drvrcdevmrsgayfsdflqqtkgkppswnvpspfslafdptkglfdyystvdevrgrrf   230
A.arachidicola_161418           drvrcdevmrsgayfsdflqqtkgkppswnvpspfslafdptkglfdyystvdevrgrrf   229
A.parasiticus_236245            -----devmrsgasfsdflqqtkgkppswnvpspfslafdptkglfdyystvdevrgrrf   221
A.novoparasiticus_191948        drvrcdevmrsgayfsdflqqtkgkppswnvpspfslafdptkglfdyystvdevrgrrf   230
                                                                *:*******. *** *****

A.flavus_36747                  dlgmggteatkplveemfdfsslpegstvvdvgggrghlsrrvsqkhphlrfivqdlpav   273
A.bombycis_5312                 dlgmggteatkplveemfdfnslpegstvvdvgggrghlsrrvsqkhsqlkfivqdlpav   285
A.pseudocaelatus_279719         dlgmggteatkplveemfdfsslpegstvvdvgggrghlsrrvsqkyphlrfivqdlpav   290
A.pseudotamarii_276010          dlgmggteatkplveemfdfsslpegstvvdigggrghlsrrvsqkyphlrfivqdlpav   290
A.nomius_129446                 dlgmggteatkplveemfdfsslpegstvvdvgggrghlsrrvsqkhphlkfivqdlpav   281
A.pseudonomius_274161           dlgmggteatkplveemfdfsslpegstvvdvgggrghlsrrvsqkhphlkfivqdlpav   281
A.sergii_195135                 dlgmggteatkplveemfdfsslpegstvvdvgggrghlsrrvsqkhphlkfivqdlpav   290
A.minisclerotigenes_217119      dlgmggteatkplveemfdfsslpegstvvdvgggrghlsrrvsqkhphlkfivqdlpav   281
A.transmontanensis_544984       dlgmggteatkplveemfdfsslpeestvvdvgggrghlsrrvsqkhphlrfivqdlpav   281
A.oryzae_5677                   dlgmggteatkplveemfdfsslpegstvvdvgggrghlsrrvsqkhphlrfivqdlpav   290
A.parvisclerotigenus_400163     dlgmggteatkplveemfdfsslpegstvvdvgggrghlsrrvsqkhphlrfivqdlpav   290
A.arachidicola_161418           dlgmggteatkplveemfdfsslpegstvvdvgggrghlsrrvsqkhphlrfivqdlpav   289
A.parasiticus_236245            dlgmggteatkplveemfdfsslpegstvvdvgggrghlsrrvsqkhphlrfivqdlpav   281
A.novoparasiticus_191948        dlgmggteatkplveemfdfsslpegsavvdvgggrghlsrrvsqkhphlrfivqdlpav   290
                                ********************.***** *:***:***************: :*:*********

A.flavus_36747                  ihgvedtdkvtmmehdirrrnpvrgadvyllrsilhdypdaacveilsnivtamdpsksr   333
A.bombycis_5312                 iqgvedtdqvtmmehdirhpnpvrgadvyllrsilhdypdaacveilsnivsamdpvksr   345
A.pseudocaelatus_279719         ihgvedtdkvtmmehdirhpnpvrgadvyflrsilhdypdaacveilsnivsamdptksr   350
A.pseudotamarii_276010          ihgvddtdkvtmmehdirhpnpvrgadvyflrsilhdypdaacveilsnivsamdptksr   350
A.nomius_129446                 ihgvedtekvtmmehdirhpnpvrgadvyllrsilhdypdaacveilsnivsamdptksr   341
A.pseudonomius_274161           ihgvedtdkvtmmehdirhpnpvrgadvyllrsilhdypdaacveilsnivsamdptksr   341
A.sergii_195135                 ihgvedtdkvtmvehdirhpnpvrgadvyllrsilhdypdaacveilsnivsamdpiksr   350
A.minisclerotigenes_217119      icgvedtdkvtmmehdirhpnpvrgadvyllrsilhdypdaacveilsnivsamdpiksr   341
A.transmontanensis_544984       ihgvedtdkvtmmehdirhqnpvrgadvyllrsilhdypdaacveilsnivsamdptksr   341
A.oryzae_5677                   ihgvedtdkvtmmehdirrpnpvrgadvyllrsilhdypdaacveilsnivtamdpsksr   350
A.parvisclerotigenus_400163     ihgvedtdkvtmmehdirrpnpvrgadvyllrsilhdypdaacveilsnivtamdpsksr   350
A.arachidicola_161418           ihgvedtdkvtmmehdirrpnpmrgadvyllrsilhdypdaacveilsnivtamdpsksr   349
A.parasiticus_236245            ihgvedtdkvtmmehdirrpnpvrgadvyllrsilhdypdaacveilsnivtamdpsksr   341
A.novoparasiticus_191948        ihgvedtdkvtmmehdirrpnpvrgadvyllrsilhdypdaacveilsnivtamdpsksr   350
                                * **:**::***:*****: **:*******:**********************:**** ***

A.flavus_36747                  illdemimpdllaqdsqrfmnqidmtvvltlngkerstkewnslitmvdnrletekiwwr   393
A.bombycis_5312                 illdemvmpdllaqdsqrfmnqidmtvvltlngkerstkewdsliatvegglktektwwr   405
A.pseudocaelatus_279719         illdemvvpdllaqdsqrfmnqidmtvvltlngkerstkewdsliatvgggleiektwwr   410
A.pseudotamarii_276010          illdemvvpdilaqdsqrfmnqidmtvvltlngkerstkewdslitavdgggleiektwwr   410
A.nomius_129446                 illdemvvpdllaqdsqrfmnqidmtvvltlngkersakewdslittvdgkletektwwr   401
A.pseudonomius_274161           illdemvvpdllaqdsqrfmnqidmtvvltlngkersakewdslittvdgkletektwwr   401
A.sergii_195135                 illdemimpdflaqhsqrfmnqidmtvvltlngkerstkewdsliktvdgrletektwwr   410
A.minisclerotigenes_217119      illdemimpdllaqdsqrfmnqidmtvvltlngkerstkewdslittvdarletektwwr   401
A.transmontanensis_544984       illdemimpdllaqdsqrfmnqidmtvvltlngkerstkewdslitavdgkletekiwwr   401
A.oryzae_5677                   illdemimpdllaqdsqrfmnqidmtvvltlngkerspkewnslitmvdgrletekiwwr   410
A.parvisclerotigenus_400163     illdemimpdllaqdsqrfmnqidmtvvltlngkerspkewnslitmvdgrletekiwwr   410
A.arachidicola_161418           illdemimpdllaqdsqrfmnqidmtvvltlngkerstkewnslittvdgrletekiwwr   409
A.parasiticus_236245            illdemimpdllaqdsqrfmnqidmtvvltlngkerstkewnslittvdgrletekiwwr   401
A.novoparasiticus_191948        illdemimpdllaqdsqrfmnqidmtvvltlngkerstkewnslittvdgrletekiwwr   410
                                ******::**:***.********************** ***:***  *   *: ** ***

A.flavus_36747                  kgeegshwgvqqlrlrk*---- 410
A.bombycis_5312                 qgeegshwgvqqlrlrgnsak* 426
A.pseudocaelatus_279719         kgeegshwgvqqlrlrgnsak* 431
A.pseudotamarii_276010          kgeegshwgvqqlrlrgnsak* 431
A.nomius_129446                 kgeegshwgvqqlrlhgslan* 422
A.pseudonomius_274161           kgeegshwgvqqlrlhgslan* 422
A.sergii_195135                 kgeegshwgvqqlrlrk*---- 427
A.minisclerotigenes_217119      kgeegshwgvqqlrlrk*---- 418
A.transmontanensis_544984       kgeegshwgvqqlrlrk*---- 418
A.oryzae_5677                   kgeegshwgvqqlrlrk*---- 427
A.parvisclerotigenus_400163     kgeegshwgvqqlrlrk*---- 427
A.arachidicola_161418           kgeegshwgvqqlrlrk*---- 426
A.parasiticus_236245            kgeegshwgvqqlrlrk*---- 418
A.novoparasiticus_191948        kgeegshwgvqqlrlrk*---- 427
                                :***************:
```

Figure B.12 Investigation of *aflP* and *aflQ*. The *aflP* gene is important for the last steps in the biosynthesis of aflatoxin, but it is missing in most of the predicted aflatoxin clusters. Here is an alignment of the aflP protein against the best hits in the other *Flavi* species.

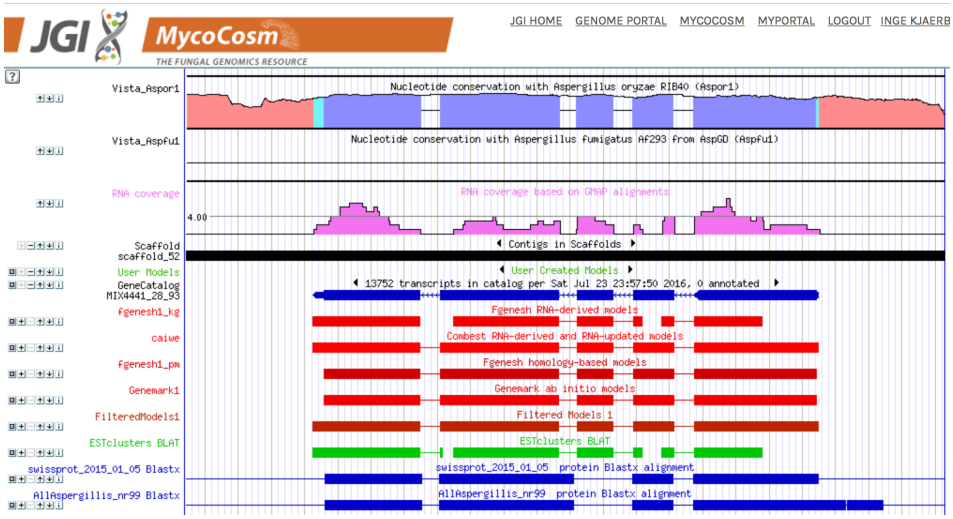


Figure B.13 Gene prediction of *aflP*. An overview of the gene prediction in the *aflP* gene in *A. paraciticus* and the RNA coverage from JGI at Mycocosm.

C Supplementary material section 4.1 – Manuscript III

Additional files for the manuscript: 'Resistance Gene Directed Genome Mining of 50 *Aspergillus* species' presented in section

Figure C.1 Common InterPro domains in secondary metabolism. Visualization of the most common InterPro annotations of secondary metabolite genes (found in more than 1000 secondary metabolite proteins) and the size of the protein families. Two horizontal lines indicate the recommended protein family size cut-offs where X_{Input} is 2 (102) and 3 (153).

Table C.1 Species used in this study, showing species name, section, and link to the JGI pages with the genomes.

Supplementary Table C.1 can be found using the following link:
https://files.dtu.dk/u/86_pamYv6WThr-lo/TableC1_SpeciesOverview.xlsx?l

Table C.2 Overview of the 72 identified putative resistance genes families and the parameters where they were identified.

Supplementary Table C.2 can be found using the following link:
https://files.dtu.dk/u/46k5NnYpOTvxqTk0/TableC2_SelectedHfam.csv?l

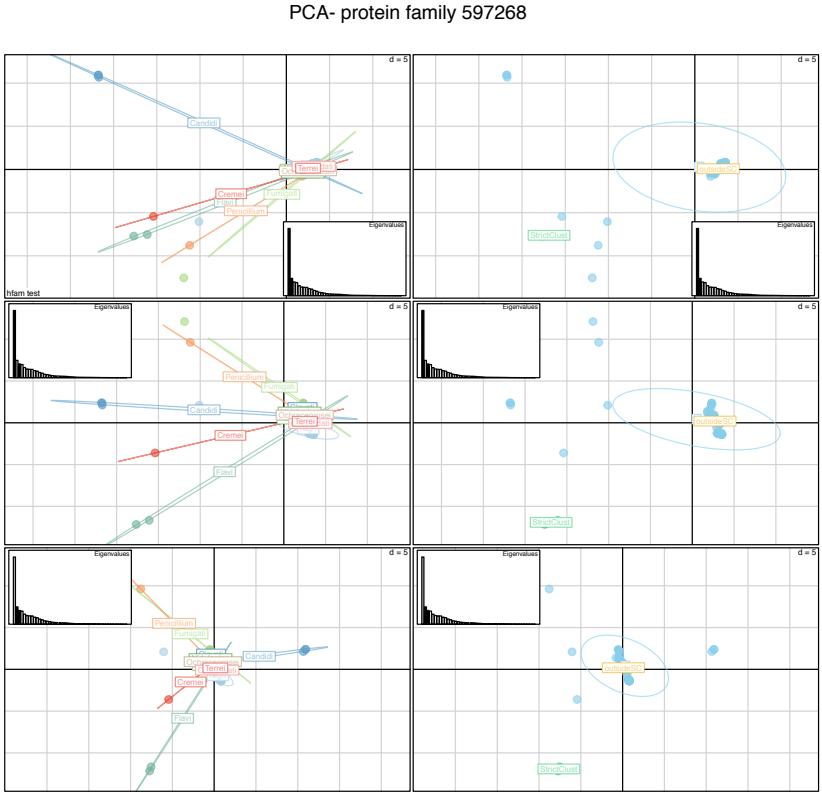


Figure C.2 Principal component analysis of the protein family 597268 containing two potential resistance genes (protein id 11595 and 32200) found in *A. oryzae* and *A. flavus*. The panels to the left are colored based on the sections the proteins belong to while the panels to the right are colored based on if the protein if found in a a selected cluster (StrictClust- green), not in a cluster (0-blue), and the homolg to the ones found in selected clusters (outsideSC-yellow).

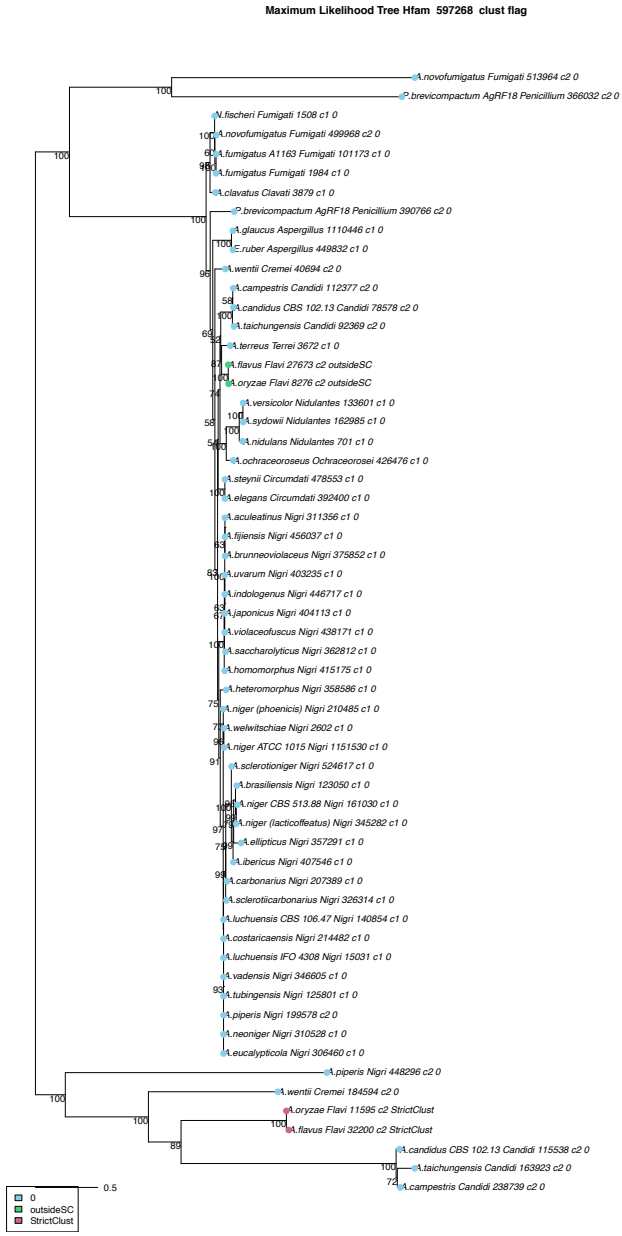


Figure C.3 Phylogenetic tree of protein family 597268 containing two potential resistance genes (protein id 11595 and 32200) found in *A. oryzae* and *A. flavus*. The node labels shows bootstraps values based on 500. The tip labels have the species name, the section, protein id, number of homologs in the species and an indication if it is found in a selected cluster (StrictClust), a predicted cluster (Clust), not in a cluster (0) and the homolg to the ones found in selected clusters (outsideSC)

D Supplementary material section 4.2 – Experimental investigation

Table D.1 Overview of used and created strains.

Species	genotype	references
A. oryzae A1560	*pyrG△, ku70△*	Aspergillus strain collection
A. oryzae A1560	*pyrG△*	Aspergillus strain collection
A. nidulans IBT29539	*argB2, pyrG89, veA1, nkuA△*	Aspergillus strain collection
A. nidulans	*argB2, pyrG89, veA1, nkuA△, IS4::T11594-11594-AflaPh3*s/h4-11596-T11596::pyrG*	This study
A. nidulans	*argB2, pyrG89, veA1, nkuA△, IS4::T11594-11594-AflaPh3/h4-11596-T11596::pyrG*	This study
A. nidulans	*argB2, pyrG89, veA1, nkuA△, IS4::pyrG*	This study
A. aculeatinus	*pyrG△*	Aspergillus strain collection
A. aculeatinus	*pyrG△, R::T11593-11593-AterPh3*s/h4-11595-T11595::pyrG*	This study
A. aculeatinus	*pyrG△, R::T11593-11593-AterPh3/h4-11595-T11595::pyrG*	This study

Table D.2 Overview of primers used in this study. U=2-deoxyuridine. The primer sequence is shown 5'-3')

Supplementary Table D.2 can be found using the following link:
https://files.dtu.dk/u/1vzZDUN82QLE_NkP/Atable_D2_primers.xlsx?l

Table D.3 Overview of used and created plasmids.

Collection	Plasmid discriptopn	Genotype	Creation
pAC478	P6_fl	USER-DR-*AflapyrG*-DR-USER	collection
pFC330	pCas9-pyrG t2	*AMA1-pyrG-Cas9* -PacI/Nt.BbvCI cassette	collection
pFC332	pCas9-hph t1	*AMA1-hph-Cas9* -PacI/Nt.BbvCI cassette	collection
pAC902	pCRISPR1-t1p	*AMA1-pAfumU3-Anid glycin tRNA-tU3-Cas9-pyrG*	collection
pAC975	CRISPR tRNA albA	*AMA1* -pyrG-*Cas9* -gRNA-fwnA	collection
pAC873	P6_fl_11594	11594_UP-DR-*AflapyrG*-DR-11594_Down	this study
pAC874	P6_fl_11595	11595_UP-DR-*AflapyrG*-DR-11595_Down	this study
pAC963	P6_fl_11594-11595	11594_UP-DR-*AflapyrG*-DR-11595_Down	this study
pFC990	pCRISPR_D11594 hph	*AMA1-hph-Cas9* -gRNA-11594s+e	this study
pFC991	pCRISPR_D11594+11595 hph	*AMA1-hph-Cas9* -gRNA-11594s11595e	this study
pAC995	pCRISPR_D11594 pyrG	*AMA1-pyrG-Cas9* -gRNA-11594s+e	this study
pFC996	pCRISPR_D11594+11595 pyrG	*AMA1-pyrG-Cas9* -gRNA-11594s11595e	this study
pAC223	pU2005-4	IS4Up-USER-DR-*AfupyrG* -DR-IS4Down	collection
pAC1560	pU2005_4_11594dualP11596*shortP	IS4Up-*11594-pAflaH3/H4*s-11596* -DR-*AfupyrG*-DR-IS4Down	this study
pAC1562	pU2005_4_11594dualP11596	IS4Up-*11594-pAflaH3/H4-11596* -DR-*AfupyrG*-DR-IS4Down	this study
	pIS1_argB	IS1Up-USER-*AorargB* -IS1Down	collection
pAC1561	pIS1_argB_11595dualP_11593	IS1Up-*11595-pAterH3/H4-11593* -*AorargB* -IS1Down	this study
	fwnA_del_HR	fwnA_Up-*pyrG*- fwnA_Down	collection

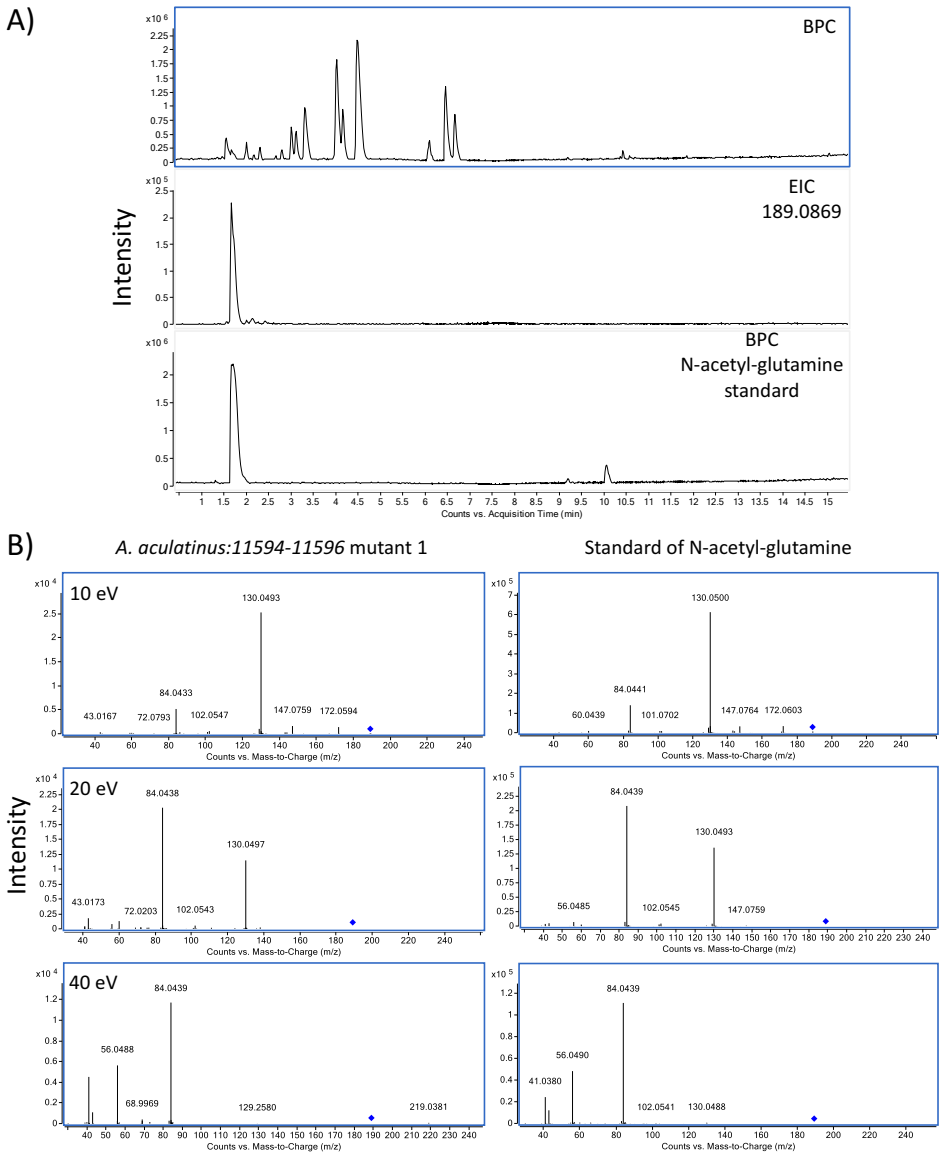


Figure D.1 Compound withm/z for $[M+H]^+$ 189.0869 and N-acetyl-glutamine standard. A) Top row: base peak chromatogram of the crude extract *A. aculatinus* Δ*pyrG*:11594-11596 mutant 1 on MM in positive ionization mode. Middle row: extracted ion chromatogram of m/z 189.0869 of *A. aculatinus* Δ*pyrG*:11594-11596 mutant 1 on MM in positive ionization mode. Bottom row: base peak chromatogram of the N-acetyl-glutamine standrad in positive ionization mode. Identity was confirmed by correct retention time. B) Tandem mass spectrometry (MS/HRMAS) spectra at 12, 20 and 40 eV of peak identified as N-acetyl-glutamine (m/z 189.0869) of *A. aculeatinus* Δ*pyrG*:11594-11596 (left) compared to the N-acetyl-glutamine standard (right).

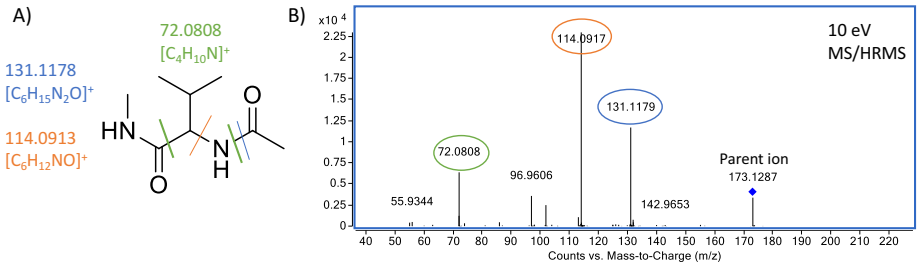


Figure D.2 A) Possible fragmentation of N-acetyl-valine methylamide corresponding to observed fragmentation ions in the MS/HRMS spectra at 10 eV B) MS/HRMS spectrum at 10 eV of compound from *A. aculatinus* Δ*pyrG*:11594-11596 with *m/z* 173.1285

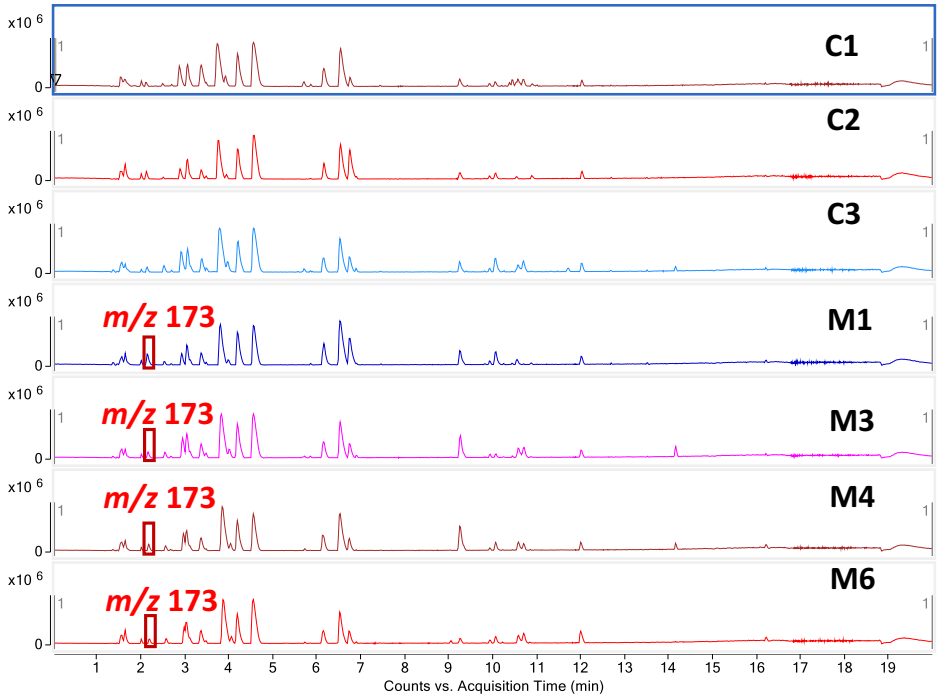


Figure D.3 Base peak chromatograms in positive ionization mode of the crude extracts of wild type *A. aculatinus* (row1-3), and four different mutants with the NRPS-like (protein ID 11594) and N-acetyltransferase (protein ID 11596) randomly inserted. Display the identification of the peak with m/z for $[M+H]^+$ 173.1285 only present in the four mutants.